10-28-2019

Study of Human Hand-Eye Coordination Using Machine Learning Techniques in a Virtual Reality Setup

Kamran Binaee
kb4000@rit.edu

Study of Human Hand-Eye Coordination Using Machine Learning Techniques in a Virtual Reality Setup

by

Kamran Binaee

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Chester F. Carlson Center for Imaging Science
College of Science
Rochester Institute of Technology

Signature of the Author _____

Accepted by _____ **10/28/2019**
Coordinator, Ph.D. Degree Program Date

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE
COLLEGE OF SCIENCE
ROCHESTER INSTITUTE OF TECHNOLOGY
ROCHESTER, NEW YORK

CERTIFICATE OF APPROVAL

Ph.D. DEGREE DISSERTATION

The Ph.D. Degree Dissertation of Kamran Binaee
has been examined and approved by the
dissertation committee as satisfactory for the
dissertation required for the
Ph.D. degree in Imaging Science

Dr. Gabriel J. Diaz, Dissertation Advisor

Dr. Reynold Bailey, External Chair

Dr. Jeff B. Pelz

Dr. Christopher Kanan

10/28/2019
Date

Study of Human Hand-Eye Coordination Using Machine Learning Techniques in a Virtual Reality Setup

by

Kamran Binaee

Submitted to the
Chester F. Carlson Center for Imaging Science
in partial fulfillment of the requirements
for the Doctor of Philosophy Degree
at the Rochester Institute of Technology

Abstract

Theories of visually guided action are characterized as closed-loop control in the presence of reliable sources of visual information, and predictive control to compensate for visuomotor delay and temporary occlusion. However, prediction is not well understood. To investigate, a series of studies was designed to characterize the role of predictive strategies in humans as they perform visually guided actions, and to guide the development of computational models that capture these strategies. During data collection, subjects were immersed in a virtual reality (VR) system and were tasked with using a paddle to intercept a virtual ball. To force subjects into a predictive mode of control, the ball was occluded or made invisible for a portion of its 3D parabolic trajectory. The subjects gaze, hand and head movements were recorded during the performance. To improve the quality of gaze estimation, new algorithms were developed for the measurement and calibration of spatial and temporal errors of an eye tracking system.

The analysis focused on the subjects gaze and hand movements reveal that, when the temporal constraints of the task did not allow the subjects to use closed-loop control, they utilized a short-term predictive strategy. Insights gained through behavioral analysis were formalized into computational models of visual prediction using machine learning techniques. In one study, LSTM recurrent neural networks were utilized to explain how information is integrated and used to guide

predictive movement of the hand and eyes. In a subsequent study, subject data was used to train an inverse reinforcement learning (IRL) model that captures the full spectrum of strategies from closed-loop to predictive control of gaze and paddle placement. A comparison of recovered reward values between occlusion and no-occlusion conditions revealed a transition from online to predictive control strategies within a single course of action. This work has shed new insights on predictive strategies that guide our eye and hand movements.

Acknowledgements

Special thanks to my advisor Dr. Gabriel J. Diaz for his guidance throughout my study

Thanks to Dr. Jeff B. Pelz and Dr. David W. Messinger for their ever-lasting support

Thanks to Dr. Christopher Kanan and Dr. Reynold Bailey for their invaluable mentorship

To my dad Ali whose presence is the warmth of my heart

To my mom Nahid and my sister Reyhaneh the stars of my life from a far distance

To Mohsen and Majid who have always been my role models

To my friends whom I have been learning and exploring the universe with

Contents

List of Figures

Chapter 1

Introduction

Vision is our most dominant sense. Research estimates that eighty-five percent of our perception, learning, cognition and activities are mediated through vision [1, 2]. But the term vision is more than just seeing an object. Vision is the process of deriving semantics, understanding concepts and creating a representation from the scene. It is a complex, evolved and developed set of functions that involve several skills. The ultimate purpose of the visual process is to arrive at an appropriate motor and/or cognitive response. An interesting feature of our visual system is that, not only most of its extremely complicated aspects are taken for granted, but also it's hard to describe it with a computational model. Creating a model that serves in numerous circumstances and for many different purposes is extremely complicated since the functionality of our perception-action system depends on the task. This model takes input visual information and produces outputs. These outputs range from very low level meaningless features such as points, lines or edges to the high level semantic interpretations such as static shapes and/or moving objects or higher level motor commands and/or cognitive perception.

One important purpose of our visual system is to guide action [2]. Being the most dominant source of our sensory input signal, the accuracy of our actions highly depend on the quality of

the visual input [3, 4]. Actions such as reaching, grasping and intercepting an object are among examples that rely upon coordination in time and space [5, 6]. The goal of this research is to study this intricate relationship between visual processing system and the motor system in a controlled environment. We first aim to answer some basic questions related to theories of perception and action by creating a naturalistic environment and designing a controlled experiment to measure human performance and present the statistics of their behavior. Since our perception-cognition system relies on the current task in hand and it doesn't have one general solution for all tasks, we need to focus on one paradigm [2, 7, 8].

In this study visually guided target interception is chosen as the main paradigm. The first two sections of this document provides background information and describes the proposed problem to be addressed. Here the fundamental theories behind perception for action are also reviewed. Furthermore, the history of eye tracking and virtual reality environments is reviewed while briefly explaining the functionality of our visual system. Also, the important considerations that need to be taken into account when modeling human visual-motor strategies are reviewed from the perspective of machine learning techniques and it builds the foundation for the follow up studies. Section three explains the requirements for system development, software and hardware details used in this study and the data analysis pipeline. In section four the results of multiple studies focused on predictive eye-hand movement strategies are presented. Finally in section five and six the author provides a conclusion and future directions for this dissertation.

1.1 Theories of Visual Perception: Ecological Vs. Constructivist

In this section the notion of visual perception from the perspective of a vision scientist is explained and the two well known theories are reviewed to provide the required background. Imagine we are looking at a scene. Intuitively the so called process of "presentation" of the object being created in our mind, is called "visual perception." There are two major theories proposed to describe the

process of perception. The *constructivist theory* that posits, during processing, the bottom-up flow of information transitions from our eyes to a simple representation of the object and layer by layer this representation gets more complex and closer to meaningful copies of the scene in our mind, until it reaches a semantic understanding. In summary, the goal of vision from the constructivist perspective is to recreate a copy or a representation of the real world in our mind [1].

On the contrary *the ecological* theory of visual perception posits that visual perception or "awareness" happens from top to bottom where we only use the visual information available in the world to find the matching solution in our mind. From the ecological perspective, the goal of visual awareness is to present you with veridical presentations of the world around us [1]. In other words, information does not only flow from our eyes to our mind/brain, but the other way around. Such that we use the "optical invariants" available in the environment to find the matching presentation in our mind [1, 7]. Gibson refers to "optical invariants" as the geometrical relationship between objects in our field of view that contains rich, sufficient and robust information [9]. As an example, when we are walking ahead, all surrounding objects in our field of view provide information about where and how fast we are moving. This rich field of vectors pointing toward the heading direction and their magnitude is proportional to our velocity. In this example, the optic flow vectors are an example of "optical invariants" that provides information about the speed of self motion during navigation. The reason that it's called invariants is because its property is robust to distortions such as height change and head rotation.

The comparison between these two theories of information flow is shown in Figure 1.1. In his review, William Warren compares the work of well known constructivist scholar David Marr with that of J.J. Gibson, the founder of the ecological framework [2]. As Warren summarizes, despite the fact that Marr's approach has significant positive influences on the field, it suffered from some fundamental assumptions [2].

Warren cites two major shortcomings of the pure constructive theory:

1) First, we know that vision takes one-third of the brain processing power with more de-

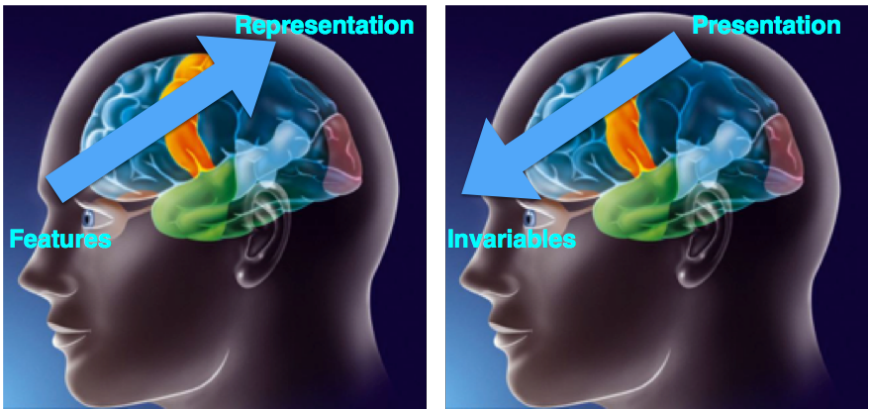


Figure 1.1: Bottom-up vs top-down visual perception theories (Image adapted from [10]).

scending than ascending connections. This might be too complicated to be reduced to a set of modular independent parts made up of bottom-up sequence of operations. Therefore an approach that doesn't take into account "top-down" constraints such as attention, the task or the intended action would not be successful [1, 2, 11].

2) Second, several studies show that the assumption that the goal of visual computation is to recover general-purpose description of a Euclidean, Newtonian world (known as "Inverse Optics"), would fail in many circumstances. It is shown in many studies that our perceptual judgments of metric Euclidean world such as depth, length, surface slant and curvature are systematically distorted. Also the problem of vision would be under-determined with no sufficient information in the two images [12, 13].

Considering the reasons mentioned above, approaching the problem of vision purely from a constructivist perspective creates a dangerous problem. This has led to models that are heavily constrained mathematically so that the algorithms converge. Without taking into account the bottom-up information, these mathematical constraints make the models too specific and narrow. As a consequence the resulting models were not robust and failed to generalize to more complex and natural conditions [2].

In conclusion it is important to note that the goal of vision is not to recover a metric, 3D physical Euclidean description of the world, as implied by the theory of inverse optics. This opens up the possibility that there is sufficient information in the input to the eye itself to specify the scene properties we do recover. Through perception, we clearly recover some aspects of the object that enable us to distinguish, identify, and interact with 3D shapes. In fact, there are many weaker geometric properties that remain invariant under affine distortions and could support the perception of objects and scenes [1]. As Gibson said, "the environment to be perceived ... is not the world of physics but the world at the level of ecology." Based on the studies mentioned in this section, the goal of vision is not to create a general-purpose representation of the world, rather, it becomes more like a collection of task-specific (top-to-bottom side by side with bottom-up) mathematical functions based on optical invariants available in the world [7, 14]. In other words the goal of vision is determined by a model that perception is uniquely determined by information considering both of the two mentioned pathways. To paraphrase Gibson, "For each perceptible entity in the world, there must be a property of stimulation, however complex, that specifies it."

In the following section we will briefly review the theories for how perception guides action, with the goal of emphasizing the difference between ecological vs. constructive explanations for the visual guidance of action.

1.2 Perception & Action

An important goal of visual perception is to guide action. In fact, we mostly use the perceived visual information to take an action such as navigating in the world [5]. Otherwise our visual perception pipeline is solely extracting environment features (or as mentioned earlier optical invariants) [9]. Now the question is: what is the framework that characterizes the transition from visual information to action? Is this a one-directional, open-loop control? Is it a closed-loop, feedback control? Are there intermediate steps being taken between extracting visual information and

taking the action? What about feed-forward pathways that can be used to predict the subsequent movement?

The categorization proposed by Warren helps us to understand, differentiate and systematically test these overlapping ideas. Warren categorizes hypotheses in the field of perception & action into three major classes [15].

- strong on-line control
- strong model-based control
- hybrid control

Here we briefly review Warren's study and provide examples for each of these categories. Most of these studies use target interception or locomotion paradigm as the most common form of visual-motor coordination task [3, 5].

1.2.1 On-Line Control of Action

In on-line control strategies, action is guided by current visual information that is available during the ongoing movement. This means there is a closed loop feedback that receives visual information, performs the action and then uses the error signal to make corrections in a closed loop way. As mentioned earlier, according to Gibson, there are several types of visual information available to specify the properties of the environment. As an example, during navigation we can determine the direction of steering based on available optic flow, which determines one's heading with respect to a target. Or, when we're visually tracking an object that is moving across the observer's field of view, the difference between target angular position on the retina and the fovea is fed to a control loop. At every time step the gaze position is proportionally adjusted based on this error signal. It is notable that, because visual-motor strategies are based entirely on online-control, these

strategies fail drastically when there is even a temporary loss of visual information, for example, due to temporary occlusion of the moving object [15, 16].

1.2.2 Internal Model based Control of Action

In model-based control, action is guided by an internal representation of the actor and the physical world. The emphasis here is that the perceived visual information passes through a "strong" internal model that produces action via a clear mathematical formulation.

Craik originally introduced the concept of internal model in his well-received book, *The nature of explanation* [17]. He proposed that the brain "imitates" a physical process by use of an "internal model of reality." As a result this method is capable of estimating external events in the physical world. Internal model control was used in engineering to tackle inherent shortcomings of a feedback loop such as delay [18]. Similarly by incorporating an internal model of the system i.e. musculoskeletal system, mobile robots improved their performance [19]. Wolpert believes that internal models are "putative neural systems that mimic physical systems outside the brain", whose "primary role is to predict the behavior of the body and the world [20]".

Unlike the on-line method that directly connects the human to the environment by use of available optical information, the internal model-based approach includes an internal model that encapsulates the world, the environment and the actor's state. In other words the actor state is monitored either through visual information or motor efference, and is used to update the world model. Thus, the term internal model refers to an inner replacement of the world and it should be robust even if the sensory input is decoupled temporarily. Thus the internal model is playing an important role for controlling the action [21].

This hypothesis is challenged when there is no visual information available such as occlusion. In this scenario we intuitively expect the on-line strategies to fail, but a "full-blooded" internal

model approach will continue to work without significant loss of performance [6, 15]. For example, it has been proposed that, when catching a ball, one can form predictions of future ball position on the basis of an internal model that includes Newton's laws, gravity and air resistance as the fixed parameters and takes ball initial position and/or velocity as input[6, 15]. If the action is derived by an internal model of ball's parabolic trajectory, when the quality of visual information is degraded (i.e., occlusion) the model performance should not deteriorate as much as a purely online form of control (i.e., ecological control).

Figure 1.2 shows the comparison between different control strategies for action. The top row shows the online control where there is a direct coupling of input visual information and action. Whereas in the middle row visual information is mediated by internal model to derive action.

Strategies that use no visual information are referred to as "offline strategies" [15, 22]. For example, in the blind walking task, blind-folded participants are able to walk successfully to a target they have previously viewed [15]. Or, during a target interception task, if the target disappears or gets occluded by another object we might use internal model of ball trajectory. Also other studies show that spatial memory is another source that controls navigation in case of no visual input available [23].

Considering the studies in this field, a major problem with internal model approach is that its parameters are rarely specified clearly. Although, both of these two hypotheses help us to understand the control of action [15], however, the online approach has a very important advantage and that is, it is succinct in terms of representation. This approach unlike the internal model doesn't require a high fidelity and detailed model to guide movement, rather it calculates simple visual cues such as optic flow, expansion rate, etc. from the available imagery. There is strong evidence that our visual system is capable of extracting the mentioned visual cues at the neural level, however, evidence is not in favor of gravity or resistance being represented in our sensory system [6, 22, 24]. On the other hand the model-based approach does not heavily rely on current visual information. Its performance is still accurate even though the visual information is unavailable. It should be

noted that a strong internal model approach should be applied to both domains, with or without reliable input visual information [15].

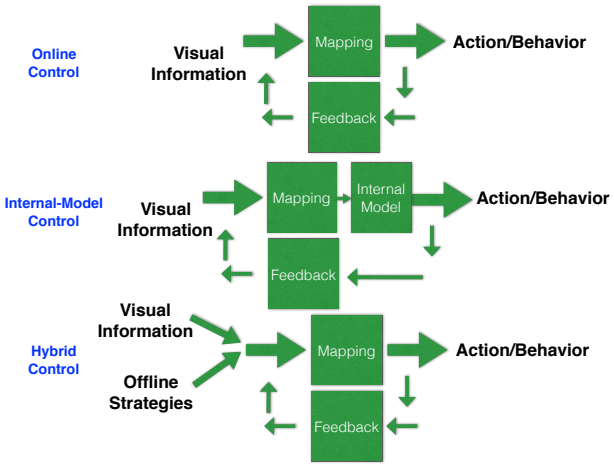


Figure 1.2: Theories of Perception and Action categorized by Warren [2]

1.2.3 Role of Prediction

Despite the issues with the model-based control, it is unclear how one can account for the inability of purely online control strategies to command action to account for behavior during temporary occlusion of the target [19]. Also there is an inherent motor-delay in our nervous system. This means, regardless of visual processing delay, from the time the appropriate signal is sent to a muscle to the time that the action is performed there is at least 150-200 milliseconds delay [25, 26]. Several studies show that with the motor-delay in the loop many fast target interceptions that humans do won't be possible [22]. Therefore there must be a feed-forward mechanism that we incorporate to overcome these problems. In fact there is evidence at the neural level that proves this feed-forward flow of information [6]. But the nature of prediction is not clear yet. There are many ways that we can model human predictive visual-motor strategies. For example, an extrapolation in time, a simple spatio-temporal mapping or an expectation prior model. The next section will

explain how these machine learning based methods would provide us with the tools to achieve this goal.

1.2.4 Hybrid Approach to Control of Action

Under special conditions, such as when the target is occluded, or the available information about the moving target's trajectory is inadequate, action may be guided in an "offline" manner. What is the exact implementation of this strategy is still an open question. The answer could be from a strong world model to a simple extrapolation of the current state of the environment. An alternative explanation is that control is mediated by a simple heuristic that falls short of an internal model, such as a simple visual-motor mapping or spatial-temporal memory. Warren calls this weak off-line strategy as a complement for on-line control without placing an undue computational burden on the visual motor system [15, 22].

Therefore according to hybrid hypothesis, on-line control is normally incorporated in the presence of reliable visual information. In case of occlusion, target disappearance or any type of visual information withdrawal, off-line heuristic or mapping strategies take place. These strategies are weak, inaccurate and short-lived, but they still provide reasonable solution to drive motor commands [15]. As an example, an internal model of ball trajectory may be expected to generalize to new conditions such as ball distance, speed and launch angle. In contrary a heuristic model is task-specific and hard to generalize to new tasks with new parameters even in the same domain.

1.3 Overview of the Thesis

The goal of this study is to investigate and model human hand-eye coordination including its important characteristics such as prediction. We chose to study hand-eye coordination in a ball catching task using a virtual reality (VR) environment. The task of catching a ball is favorable because its performance is within the ability of most healthy adults, and it has previously been the

subject of empirical study. It includes the visual information processing for the coordination of action - the hand movement. We use a virtual reality environment because it allows us to provide a naturalistic environment for the subjects along with control of experiment parameters. Furthermore, we use a combination of eye tracking and motion capture systems because it allows us to record subjects gaze, head and hand movements. Once we have collected the behavioral data of several subjects attempting to catch virtual balls, the results of their performance will be presented, and considered as evidence for or against the proposed hypothesis. The behavioral findings will also be the basis for computational, machine-learning models that test competing theories of visual and motor control. First, based on the collected dataset we create a supervised model that mimics human visual-motor performance and we present this model as an evidence for hybrid control theory of perception and action. This supervised model also provides a solution for perception-action loop that can be tested systematically and it also captures prediction. By use of visualization and statistical analysis we determine the role of individual sources of visual information on performed action. This method will also help us to determine how long our predictive strategies are valid in time. Finally we use an inverse reinforcement learning approach that allows us to characterize human performance using action-reward mechanism. The second modeling scheme will help us to address the hypothesis related to on-line vs. predictive control in a single representation approach. Since the reward function for a reinforcement learner agent presents the underlying strategies that produce the behavior, we use this tool to visualize when human subjects make the transition between on-line to prediction during the course of an action. The key component here is to define constraints on the agent so that the model captures human behavior in a realistic way.

Chapter 2

Background

In this section the main features of human visual system are briefly reviewed to help understanding the behavioral findings that we will discuss in the results section. Although physiology of our eye can be broken down into several detailed sub-systems, our purposes require only a review of the components involved in the extraction of information from the world.

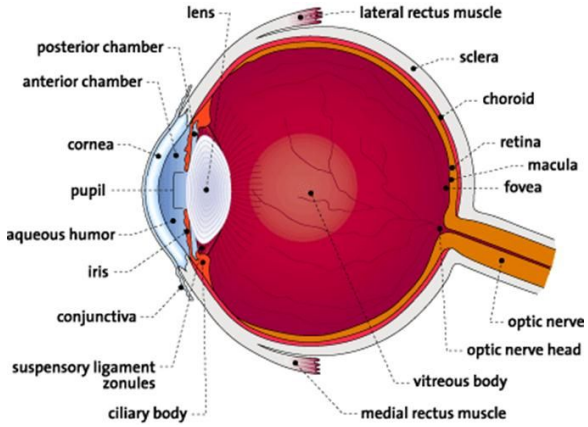


Figure 2.1: Anatomy of human eye. Image courtsey of [27]

2.1 Human Visual System

It is no exaggeration to say human eye is the most fascinating optical system. It is an extremely efficient imaging system that transforms light into electrical signal using an adaptive compression method. The pupil is the aperture of this imaging system, its size is adaptively tuned to the ambient light. Light passes through the cornea, which has approximately two-thirds of the eye's total optical power, and then passes through the lens. Lens thickness can be modified by stretching or releasing the ciliary muscles, and this determines its contribution to optical power, as well as the radius of focused light upon the retina. When we change our focus point from a far distance to a near distance object, the optical power of the lens changes accordingly so that the focal plane falls on the retina as shown in Figure 2.1.

The retina is the imaging plane for this adaptive optical system, while containing all the photoreceptors. There are two types of photoreceptors, rods and cones. Rods are mostly responsible for low light vision (scotopic) while cones are responsible for daylight color vision (photopic). Rods the more sensitive type of photoreceptors, while cones are important for providing high acuity vision. In the center of the retina there is an important region that occupies 2-3 degrees of visual angle called the fovea. The importance of this region is that it has the smallest cone photoreceptors and is almost free of rods. There are 120 million photoreceptors as shown in Figure 2.2, and they are distributed unevenly throughout the retina. There are more photoreceptors in the periphery with less density compared to the central field of view, and those further in the periphery are larger in size. This also means that once we move away from the fovea our visual acuity drops drastically. In other words, if we are fixating at an object in a scene in front of us, the center of the scene (2-3 degrees) projects to the fovea on the retina and the rest of the scene is projected on the periphery. We have our highest visual acuity at the fovea and it decays as we go further toward eccentricity. That's why we need to make eye movements and fixate on the parts of the scene where we need detail visual information. In conclusion, since we have only a tiny portion

of our visual field available for us in high quality, we need to make eye movements to bring the objects of interest on our fovea (i.e. to foveate the target). In the following subsection we review eye tracking methods and describe different types of eye movements.

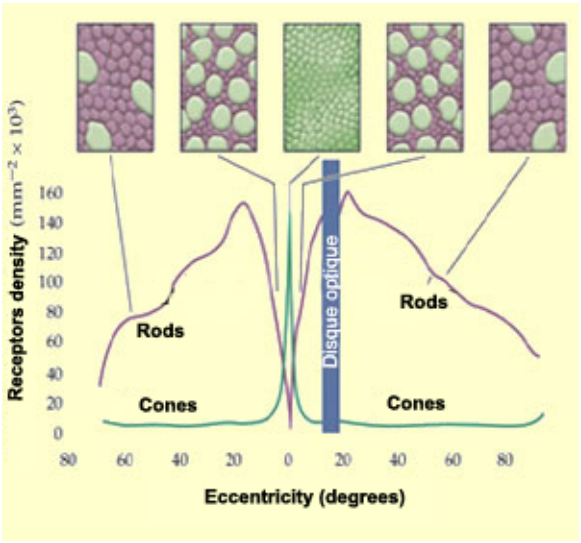


Figure 2.2: Distribution of Cones and Rods throughout the retina. Image courtesy of[28]

2.2 Study of Eye Movements Using Eye Tracking

Since the focus of this dissertation is the coordination between perceived visual information and performed action, it is important to provide a general background about different methods of eye tracking. The goal here is to introduce eye tracking techniques from a general perspective. Further details on different techniques and the history of eye tracking can be found in [29, 30].

The question of how/why/when do we move our eyes has been of interest since early 19th century. In 1908 Edmund Huey designed one of the very first eye trackers using a contact lens like device with a hole for the pupil. Later on Yarbus contributed to this field through the invention of several new apparatuses for the study of eye movements. In his influential book *Eye Movements and Vision* he wrote about the relationship between fixation and regions of interest and proposed important questions to the psychology community [31]. Prior to Yarbus, Guy T. Buswell used light beams which were reflected on reader's eyes and recorded them on film. Buswell's research, for the first time, revealed important findings about our eye movements while reading a text. There are different classes of eye tracking techniques. Here we only explain two of these methods : 1) electro-oculography, 2) video oculography.

2.2.1 Electro-Oculography

Electro-oculography is based on a sensor that detects the difference between electrical potential of the front vs. rear of the eye. In this method sensors are connected on the skin to measure the change in electric field that is created when the eyes rotate. Although this invasive method is noisy and subject to drift in the received signal, its advantage is that it can record eye movements when our eyes are closed [29].

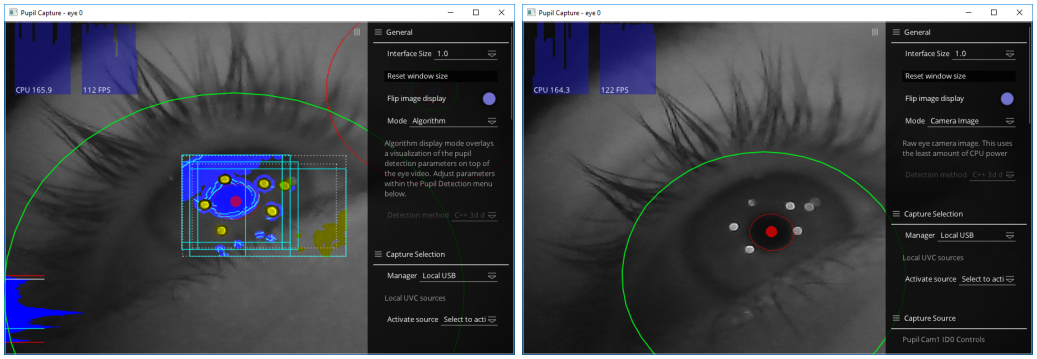
2.2.2 Video Oculography

Video oculography is based on an image of the eye and use of computer vision/image processing methods to find the position of the pupil and the cornea. The majority of the current eye tracking methods used in consumer electronic devices are video based systems. These methods use one or multiple cameras alongside with visible light or infrared light to illuminate the surface of the eye. There are multiple reflections of the light off of the surface of the eye. By using an image processing algorithm a gaze vector emanating from the pupil is calculated. The details of the process of calculating the gaze vector can vary significantly depending on the hardware, optics and software specs of a video based eye tracking system. One can simplify the steps as follows: 1) capture image(s) of the eye, 2) apply an image classification technique in order to find the position of different key elements of the image (cornea reflection, pupil center, etc.) 3) Use 2D or 3D geometry of the image(s) and the eye to calculate the gaze vector corresponding to the video frame.

There are desktop and portable versions of video based eye tracking systems. Some desktop based eye trackers have temporal resolution of up to 1000 Hz (Eye Link Inc. [32]).The portable eye trackers are more used for less constrained studies, where the subject wears the device and freely moves his or her eye and head. In addition, there are desktop eye tracking devices that allow a certain range of head translation and rotation with varying range of accuracy (Tobbii 4C tracker [32]]).

Recent algorithms use a surface model of the eye to extract the 3D gaze vector by illuminating the eye using a group of IR LEDs. As it is shown in Figure 2.3, first the pupil and the reflection of the IR LEDs are detected so that the algorithm can find an ellipse fit to the iris. Once the pupil and the iris are detected the eye surface model provides an estimate of the eye ball shown in green in Figure 2.3b. Finally, it generates a vector estimate of gaze direction emanating from the center of the fovea to the pupil centroid. These algorithms show robust performance and up to 120 Hz

refresh rate in both portable and desktop setup.



(a) (b)

(c)

Figure 2.3: View of the eye from the eye tracking camera. (a) The image processing algorithm detects the IR LED reflections and fits an ellipse to the Iris, (b) five IR LED reflections are shown on the surface of the eye and the pupil being color coded by red (c) generated 3D gaze vector and the proposed sphere fit to the eye ball (courtesy of Pupil Labs Inc.)

2.3 Virtual Reality in Research

When studying visually guided control of action, it is very important for the experimenter to have full control on experiment parameters and sometimes this comes with constraints on the condition in which the subject would perform the task. For example, in order to study motion perception for the purpose of target interception, 2D targets were presented to the subjects and their task was to predict whether the moving circle would hit a 2D line both presented on a 2D screen. A desktop eye tracker recorded subjects gaze movements and a button press was used to register their response [33]. There are several other influential studies that adopted head-fixed desktop eye tracking system and revealed important findings about our visual system and eye movements. One important characteristic of a head-fixed desktop eye tracking system is that the viewer's position relative to the screen (stimulus) is fixed. This provides accurate measurements of stimulus physical properties such as angular position and velocity.

Although studies based on head-fixed eye trackers help us better understand the mechanisms underlying visual perception and provide us with fundamental knowledge about our visual-motor processing, however some researchers would doubt whether these results are generalizable to a more realistic condition with free movements of the head and the hands. For example when someone is playing soccer and trying to predict the interception point, what type of additional source of information could they be using while moving their head and body freely in 3D space. Therefore, there is always an aspect of natural behavior that is being missed and could create biases in results or changes in behavior.

In real world, we hit the tennis ball while moving our head and also running freely in different directions. In order to run experiments in a real world condition there are tedious and precise calibration and measurement procedures required to make sure the physical properties of the visual stimulus such as size, position, velocity ,etc. are presented to the subject as intended by the experimenter. An inspiring study presented by McKinney et al investigates predictive eye movements

for squash players while they were wearing portable eye tracker [34]. This study explains the pre-processing steps required in order to analyze the eye tracking data with respect to sometimes noisy low quality image of the scene camera.

The recent developments in Virtual Reality technology has made these devices more ergonomically pleasant for users and opens an opportunity for researchers in the field of vision science. Since the VR content is pre-computed, the geometry of the virtual entities are known. Since the position of real and virtual objects are precisely calculated by the physics engine and stored during the data collection, there's no need for post-hoc recalculations of the stimulus. Unless for the purpose of spatial or temporal calibration of the VR system. Another important added value for use of VR in behavioral studies is the expedited time for prototyping and redesigning the experiment parameters which would be much harder in a real world data collection setup.

This project while being inspired by previous studies in naturalistic environments [3, 35, 36] uses a VR system in order to study eye movements in a naturalistic environment while presenting calibration techniques to make sure the accuracy of measurements and the stimulus remains valid. To provide a background, in the following subsection a brief overview of other studies that used a VR setup is presented.

2.3.1 Behavioral Study in VR

Behavioral studies that use VR as their experiment setup can be categorized into few different classes. Visually guided navigation, object and shape detection, reaching and grasping and visually guided target interception. Most of these studies use VR to first present the stimuli in a more realistic environment and/or record subject's behavior in a more naturalistic setup that provides head free eye movements and also hand movements. Furthermore, the parameters of the experiment are controlled by the experimenter which is not available in a real world scenario. The extent to which the subject feels being immersed depends on the type of VR technology used and the

need of the paradigm. For instance, in a reaching experiment where the subject is instructed to tap on a moving target on the screen, it is more important that the system has very low latency with respect to user's interaction than it is to have a very realistic appearance. In contrary, in a fully immersed VR setup that uses head mounted display (HMD) there are many parameters that can destroy the feeling of being immersed in the virtual environment. Optical aberrations introduced by the display, limited field of view and any spatial/temporal error in the system would cause significant biases on subject's perception and hence their motor response. Therefore in many studies that use immersive VR environments rely upon a meticulous process of spatial and temporal calibration [37–43].

2.3.2 Sources of Perceptual Inaccuracy in VR

To reduce the effect of spatial or temporal misalignments of the VR system on the results of a behavioral study, all aspects of a VR experiment need to be calibrated. This includes spatial and temporal calibration. Several studies that use motion capture system undergo a rigorous calibration procedure to make sure the recorded position of the markers are within the range of required accuracy. These systems either use an *active* marker technology where the markers that are being tracked are actively emitting visible/non-visible light or they use *passive* markers where the markers are made up of a retro reflective material. Both of these two types of motion tracking systems use multiple cameras (or sensor arrays) plus multi-view geometry to estimate the 3D position of markers within the tracking volume. During a typical calibration process system uses predefined reference markers to minimize the spatial error along side defining the origin of the capture volume [44].

Temporal latency is one of the most challenging problems in a VR system. Especially when we're dealing with human perception and action, if not compensated for it can jeopardize the validity of the results. Unlike spatial error, temporal inaccuracies are more challenging to be

corrected for especially during run time. Therefore, system latency has to be taken into account during experiment design and system development. Typically, behavioral studies in VR measure and report their so called *end to end latency* in order to make sure the apparatus did not create negative perceptual effects on the subjects [45–47].

2.3.3 Eye Tracking in VR

Eye tracking in an HMD based VR system is challenging due to three major reasons. Mechanical mount, display and slippage makes eye tracking more challenging compared to a portable eye tracking setup. Considering the size of a light head mounted display it is a challenging problem to find the optimum location for the built in eye tracking hardware.

Eye tracking camera needs to have a good image of the eye while it is being illuminated by IR LEDs. This makes the mechanical design challenging so that it's not occluding the display. SMI uses single sided mirror to overcome this problem [48]. Although major VR companies are trying to improve their display technology, but distortion is inevitable due to the fact that most of the head mounted devices are designed to be for consumer level products. Therefore the quality of the optics and the display is far from perfect in commonly used systems in the market. Since the presented stimulus is distorted by the display and lens, the eye tracking calibration procedure needs to take this effect into account.

Last but not the least, even with a perfectly calibrated eye tracker in an HMD VR system, since the head is free to move there's always a chance of the HMD physically being shifted on subject's face. This will cause a shift in eye image, hence an error in the reported eye tracking data. Therefore for a physically demanding experiment there needs to be repeated measures of calibration to make sure the validity of the eye tracking data throughout the time[48].

2.4 Modeling Techniques

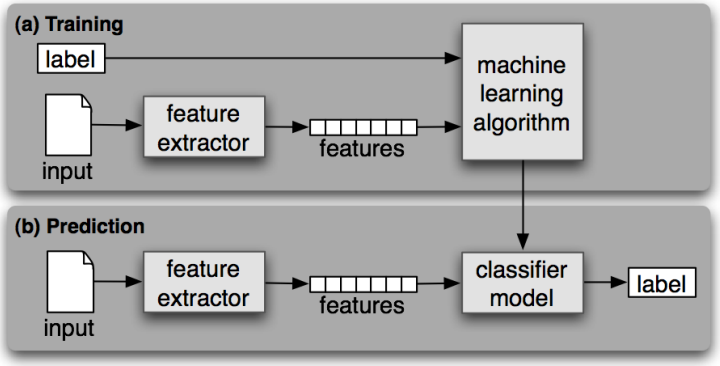
There are several modeling techniques that can be applied to human perception-action study. The rich literature of behavioral study covers wide range of models from Bayesian statistics to classical dynamic control systems. In this section a brief overview of machine learning techniques is presented. There are two major classes of machine learning that fit the goal of current research. Therefore it is important to provide a general perspective toward different goals of machine learning techniques to explain the rationale behind the choice of methods used in this study.

2.4.1 Overview of Machine Learning methods

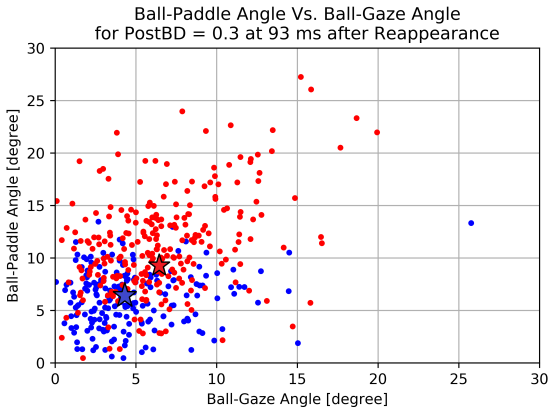
Machine learning is a science of data analysis, interpretation and classification that gives "computers the ability to learn without being explicitly programmed" according to Arthur Samuel [50]. This field of computer science has evolved from pattern recognition and computational learning theory in artificial intelligence. Which is exploring the data, finding patterns, representations and models to be able to predict the interpretation of an unseen data sample or generally make decisions. These methods allow machines or algorithms to learn these representations and even explore newly added sample data to their decision base. There are numerous applications of machine learning these days such as computer vision, image classification, robotics, data mining, etc. There are three major classes of machine learning methods: 1) supervised learning, 2) unsupervised learning and 3) reinforcement learning [49, 51, 52].

Supervised Learning:

In supervised learning the key component is that the data is available with its interpretation or so called "label." For example, in an image classification problem, most of the datasets have images with their corresponding labels. The goal of a supervised image classification algorithm is to predict the label of an image after being trained on several instances of image-label pairs. Using



(a)



(b)

Figure 2.4: (a) Flow chart of a supervised machine learning technique (diagram courtesy of [49]) and (b) Results for an unsupervised data clustering on a sample data from the VR ball catching experiment. The blue and red stars show the cluster (class) centers for the success and failure classes accordingly.

mathematical notation, in a supervised machine learning technique we have input x and output y available and the goal is to find function F that relates input and output as follows:

$y = F(x);$ where x and y are known and F is unknown

Figure 2.4a is a flow chart representing the major steps involved both training and testing a supervised model for image classification. During the training phase the model parameters are optimized by feeding the training instances and minimizing the output error with respect to the desired output (label). During testing, the model is used to generate output based on the optimized parameters, in other words the model is used to predict the best label for the input image.

Unsupervised Learning

The goal of an unsupervised learning algorithm is to find the patterns in the data so that the data can be interpreted or represented in a more meaningful way. More precisely as humans we're always interested to infer information from the data and this inference highly depends on the representation of the data. Data clustering is a perfect example of an unsupervised algorithm. Where the algorithm groups together the "close" data points so that it conveys an overall understanding of how and in what directions the data is distributed/clustered. Using the same mathematical notation here we only have x and the goal is to find F which is a better and more intuitive representation of the original data:

$F(x);$ where x is known and F is unknown

Figure 2.4b shows an unsupervised data clustering example where the algorithm represents the data using three cluster centers and cluster parameters. In other words the unknown 2D data points are categorized into three different classes based on their distance and distribution.

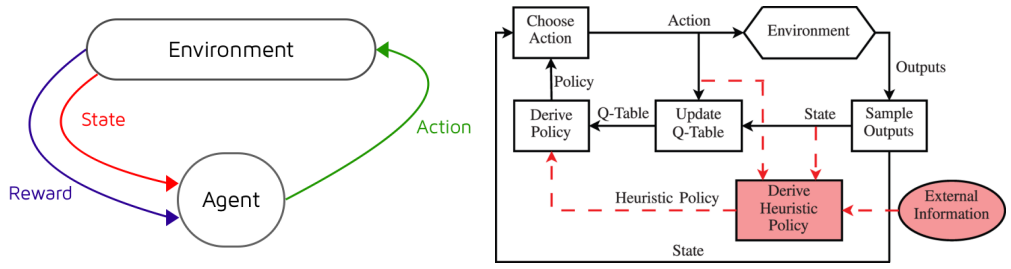
Reinforcement Learning

The goal of reinforcement learning method is to allow the algorithm to decide what label or representation is the best for the data. Reinforcement learning can be considered both a supervised and unsupervised class of machine learning. In a reinforcement learning technique the algorithm (the agent) interacts with the environment by taking an action, based on this interaction it observes its own state and the reward (reinforcement signal) that it receives from the environment [53, 54]. By repeating this so called exploration-exploitation loop several times, the agent gradually learns what is called an optimum policy. The learned optimum policy through experiencing different combination of states and actions, guarantees the maximum long term reward for the agent. A reinforcement learner agent instead of learning pairs of desired inputs-outputs, learns the optimum policy (actions) that maximizes the predefined reward function. Using the previous mathematical notation, a reinforcement learning technique could be described as the process of observing several input states x and their corresponding reward value $Z(x)$ and learn the optimum policy F that generates the best output y at every state. Note that, actions that are the means of moving between states could be included in x as state-action pair [53].

$y = F(x) \quad \& \quad Z(x); \qquad$ where x and $Z(x)$ are known and F is unknown

As shown in Figure 2.5 the agent can interact with the environment to find the optimum policy based on exploration & exploitation.

In this study we use two different classes of machine learning techniques for two reasons. First we develop a supervised learning model that is trained to reproduce subject's behavior in terms of



(a) (b)

Figure 2.5: Reinforcement learning agent and its observation-exploitation loop to learn optimum policy. Images courtesy of [53]

gaze, head and hand movements. Once we show that the model is reproducing human eye-hand movements with a reasonable range of error, we study the model to learn human visual-motor strategies. This approach allows us to understand the underlying control mechanisms through a surrogate model. Using a systematic approach we test the model to draw parallel lines between how the simplified model operates to understand how the brain would work.

Second we use reinforcement learning (RL) framework that will be explained in details in chapter 4 to find subjects underlying strategy that produced the state-action pairs. More specifically we take the inverse path of an RL model. Since we don't know the rewarding mechanism that subjects were using when performing the actions, the inverse reinforcement learning paradigm recovers the reward function based on a set of observed state-action pairs. This is also referred to as immitation learning where by using some assumptions the reward function is estimated so that it produces the similar behavior as observed by the humans [55].

2.5 Considerations for Modeling Human Hand-eye Coordination

Consider a ball catching paradigm where the subject is instructed to intercept with a flying ball in his field of view. The ball launches from a certain distance, flies in the air following the physics of gravity and passes by the subject. Intuitively we expect the subject to track the ball with a combination of head + eye movements, initiate the hand movement and reach to catch the ball. To be able to investigate the contribution of online or internal model based control, let's assume that we can make the ball disappear for a short period of a time, called "blank duration." The hypothesis is that subjects mostly rely on online control strategies as long as the ball is visible and use "some type of" offline or internal model or some simple heuristics to guide their hand movement during the blank prior to attempted catch.

To determine what is the nature of these strategies and how long they are valid in time we propose to take two important steps. First create a naturalistic environment and record human hand, head and eye movement data and second to create a model that can be systematically investigated. By using the human data set we can get an understanding of what is the common pattern of behavior and based on that we can probe, modify and recreate different models to understand and visualize these strategies. We propose to use a virtual reality setup to create a naturalistic environment that would make our modeling results compatible to real world situation. Also, because we are using a head mounted display to render the virtual content we can investigate the role of visual information available in the image. In this section we review different aspects of human behavior during a visually guided target interception task. We expect the proposed model to capture these behaviors or at least provide a systematic way to probe the presence of these characteristics.

2.5.1 From Information Space to Action State

As mentioned earlier, it is very important to study the mechanisms underlying perception and also organization of action. Our perception system receives a large number of sensory inputs in various

forms and based on that it produces the appropriate action (See Figure 2.6). In our paradigm, when the ball is moving in our field of view studies show that one can couple hand position to the perceived passing distance of the ball, which is instantaneously specified throughout the ball's approach in the form of the optical variable $\dot{\theta}/\dot{\phi}$, where $\dot{\theta}$ is ball lateral velocity and $\dot{\phi}$ is ball expansion rate [56]. The proposed model should allow us to probe into these relationships. Also it should be noted that not all the input sensory information is visual. The human vestibular system is sensitive to orientation and acceleration during movement, and this provides important information that can be used to drive head and eye movements [57]. Furthermore our brain stores a copy of signals sent to muscles that is called "efference copy." This sensory signal provides non-visual information about the position of our hand. In conclusion as it is shown in Figure 2.6 there are visual and non visual sources of information that contributes to motor action when performing a hand, gaze or head movement. We can formalize this relationship as follows:

$y_j = F(x_1, x_2, ..., x_i)$ (1)

where j iterates on different actions and i iterates on different input sensory information (such as optical variables, proprioceptive information, etc) We expect the model to allow us to investigate the relationship between the input sensory information and output action in a systematic way. It is shown in several previous studies that we use different sources of information at different times. In example when we are intercepting with an object moving in our field of view we use target lateral velocity and expansion rate to plan our hand movement timing[56, 58]. Also not all the visual information available for us are taken into account equally at different times. Our visual perception system efficiently chooses source of information based on reliability and usefulness for the task in hand[56, 58–63].

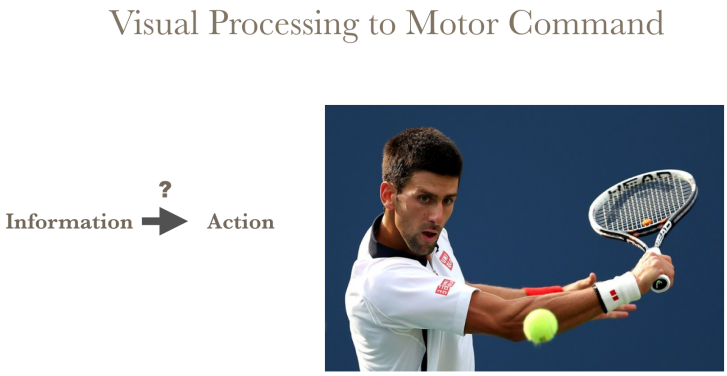


Figure 2.6: From information space to action state. Images adapted from

2.5.2 Predictive Model

One of the key features of human perception-action system that shows itself in many different shapes is that we rely on "predictive strategies" when we're performing an action. The major reason for the need for predictive strategy is the well-known motor delay. This delay includes the time that the visual information is transmitted from photoreceptors to visual cortex and also the time that it takes for a motor command to reach our muscles to make the planned movement. It is important to study the accuracy of vision and motor control in the presence of established visual and motor delays of 150-200 ms [26]. If not compensated for, these delays would cause a tracking gaze of a quickly moving target to lag behind the target, and an attempted manual interception to be poorly timed with the target's arrival. Despite evidence suggesting the presence of predictive strategies the nature of prediction is not well studied yet. Therefore we expect the proposed study would provide more detailed understanding of how the prediction mechanism is employed when we're trying to catch a target. Figure 2.7 shows one possible solution for our predictive strategies. Considering a parabolic trajectory of a target moving in our field of view our visual system might

extract information about its movement during a so called *integration time* and use that information to predict an action that is going to be taken in the future. This is basically a mapping between previously available sources of information and future action in time. This framework can be applied to model hand-eye coordination both from the perspective of online control and internal model.

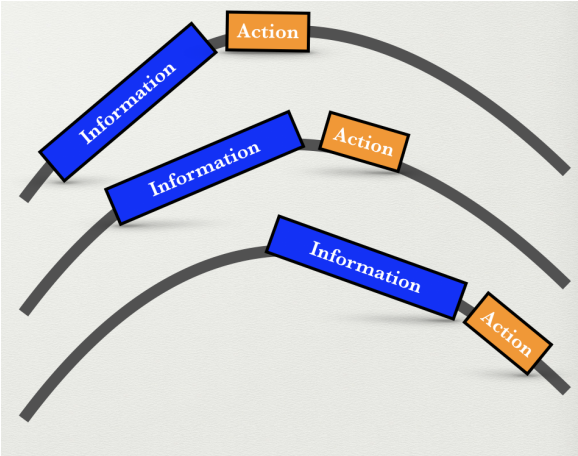


Figure 2.7: A model that integrates information in time and predicts future state pictured over a hypothetical ball trajectory, the integration duration and prediction window are shown in blue and orange respectively

Chapter 3

Methodology

In this chapter the data collection procedure using motion capture system, VR head mounted display (HMD) and built-in eye tracker is explained. In order to design an experiment to study human behavior in a VR setup it is important to measure the accuracy of each VR sub-system to make sure that the perceived VR stimulus is as close as possible to the real world stimulus. Furthermore, our data recording and processing pipeline is explained so that the reader could reproduce the results. First, the system hardware specifications are explained and then the software used for graphics rendering and data recording is discussed and finally the data processing and analysis is briefly explained.

3.1 System Hardware, Motion Capture and Head Mounted Display

In our experiments, we use *PhaseSpace Inc.* [64] motion capture system to record movements of certain objects. i.e. the head mounted display is being tracked using a set of active LED markers positioned precisely on it and this mode is called rigid-body tracking. The motion capture system is composed of 14 PhaseSpace cameras placed in the ceiling. These cameras are horizontal and

vertical linear detectors that are sensitive to the red LED lights on generated by the active marker modules. The active markers have individual IDs (A, B, C, D & E) that corresponds to the blinking frequency and shift controlled by the server. To be able to get position and orientation of for example the head mounted display, the marker IDs on the device need to be defined for the server first. This way the server would be able to detect the LED markers according to the predefined list. After calibrating the motion capture system using the calibration wand and defining the world coordinate system axis and origin, then the server starts streaming the 3D position and orientation (using quaternions) of any regid body upon clients request. This is shown in Figure 3.2.

Stimuli were delivered by an Intel i7-based PC with an NVIDIA GTX 690 connected to the Oculus Rift DK2 head mounted display at 75 Hz, and an NVIDIA GTX 760 connected to the experimenter's desktop display. The computer ran Windows 7, and the virtual environment was rendered using the Vizard Virtual Reality toolkit by Worldviz(Vizard 5.1, 2015). Physics were simulated using the OpenODE physics engine (Open Dynamics Engine, 2009) so that ball trajectories matched those expected within a real-world environment in the absence of wind resistance (Figure 3.1). Collisions between the ball and the paddle or other surfaces was also detected using the physics engine. To improve the fidelity of the physical simulation and collision detection, the physics engine was updated 10 times between frames (e.g. 750 Hz).

The Oculus Rift HMD has an approximate diagonal field of view of 100°, and an approximate angular resolution of 10-15 pixels per degree dependent upon the position of the eye inside the head mounted display [48]. Head and paddle position/orientation were sampled and recorded at 75 Hz using a 14 camera *Phasespace* X2 motion capture system, with a measured latency of between the time a sensed movement would be reflected on the display of less than 30 ms. Eye movements were recorded with an *SMI* binocular eye tracker (v0.5 beta, 2015) sampled at 75 Hz. A post-hoc correction was applied, as described in [48], to correct for helmet slippage and other sources of spatial error in the eye tracking signal. The average eye tracking accuracy after calibration and correction was 0.53° for the central visual field ($FOV < 10^{\circ}$) and 2.51° in the

periphery ($10^\circ < FOV < 30^\circ$). This method is described in more detail in chapter 4.

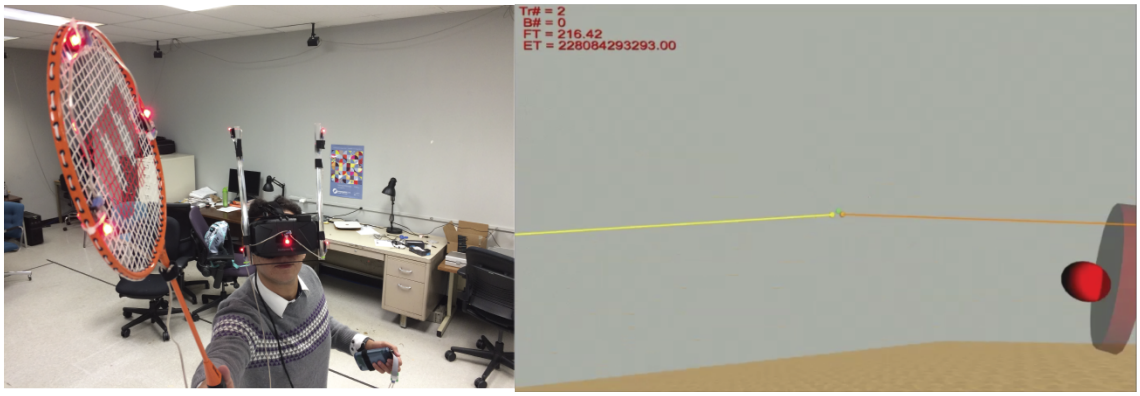


Figure 3.1: **A.** Subjects wore an Oculus DK2 with integrated SMI eyetracker with a sampling rate of 60 Hz. **B.** The experimenter's desktop view of what the subject saw inside the helmet. The lines receding in depth represent the left and right gaze vectors, and the text in the upper left encodes trial number, block number, and time-stamps. The red disc represents the face of the paddle, on which can also be seen the red virtual ball.

3.2 System Software, Graphics and Eye Tracking Data Recording

As mentioned above, the VR graphical environment was created in Vizard 5 software provided by *WorldViz* using Python. The software developed for the experiment has several sub-modules each of them responsible for a certain task. The refresh rate is set to 75 Hz and each of these modules gets updated sequentially. The motion capture interface calls the server and updates the rigid body position and orientation buffers, while the physics engine runs the collision detection routines accordingly. Once the position and orientation of all the nodes are updated the eye tracking server is called for streaming the last updated gaze data structure. The last operation is writing out the data into a structured text file which will be saved and stored onto the disk at the end of the experiment. The experimenter can also interrupt the ball catching experiment by pressing certain keys. For example, by pressing "c" the eye tracking calibration routine runs where the stimulus

appear at certain location and updated based on consecutive key press.

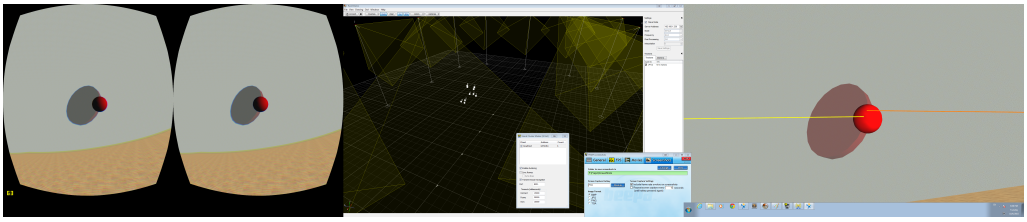


Figure 3.2: The view inside head mounted display, the motion capture software markers for paddle and the view on experimenter monitor that also shows the rendered right and left gaze vector.

3.3 Parsing the Data: From Structured Text to Pandas Data Frames

As mentioned before, the recorded data for each participant is saved into a folder named by the date and time of experiment. This folder beside the experiment data (i.e. the structured text) contains the meta-data regarding the parameters of the experiment and system hardware configuration for possible future use. The data processing starts with parsing the structured text file into pandas data frame. Each subject's data is parsed into four data frames. First, the structured text is parsed into raw data frame that is an efficient tool to handle large data sets and run fast parallel computations. The raw data frame rows are total number of frames of data recorded during the ball catching experiment. The columns of raw data frame are different physical measurements or condition flags related to the experiment. The frames of the data recorded during eye tracking calibration session is stored separately called calibration data frame. This data frame is used to calculate the calibration matrices required for compensating the eye tracking error. After pre-processing the raw data, noise and outlier removal using pandas built-in functions, gemoterical calculations of the gaze vectors i.e., gaze-in-world, the processed data frame is generated that has fewer columns yet a more clear parameters of data for further statistical analysis. Finally, the output of a specific analysis is stored into so called trial information data frame. The number of rows in the trial info

data frame is the number of trials. This allowed us to store several metrics and analysis results for each trial individually. The value of the data collection pipeline used in this study becomes more clear when we consider the total data collection, data parsing and analysis time. As an example, for a 150 trials ball catching experiment, from the time that the subject is prepared to start the experiment with the instructions, to the time that subjects initial statistical analysis is calculated, is less than 45 minutes. This allowed us fine-tuning the study using faster and more efficient iterations.

```
In [3]: rawDataFrame.head(5)
```

Out[3]:

	frameNumber	IOD	IPD	ballFinalPos			ballInitialPos		
				X	Y	Z	X	Y	Z
0	10801	60.5604	62.0453	0.001897723	1.800452	0	-4.932819	0.9686899	20
1	10802	60.5604	62.0453	0.001897723	1.800452	0	-4.932819	0.9686899	20
2	10803	60.5605	62.071	0.001897723	1.800452	0	-4.932819	0.9686899	20
3	10804	60.5606	62.0588	0.001897723	1.800452	0	-4.932819	0.9686899	20
4	10805	60.5611	62.0572	0.001897723	1.800452	0	-4.932819	0.9686899	20

```
In [5]: trialInfoDataFrame.head(5)
```

Out[5]:

trialNum	ballCaughtFr	ballCaughtQ	blankDur	postBlankDur	preBlankDur	firstFrame	lastFrame
0	110	True	0.5	0.4	0.6	0	299
1	450	True	0.5	0.5	1.0	300	599
2	712	True	0.5	0.4	0.6	600	899
3	NaN	False	0.5	0.3	1.0	900	1199
4	1301	True	0.5	0.3	0.6	1200	1500

Figure 3.3: The raw and trial information data frames as stored during data processing. The number of rows for raw data frame is equal to the number of frames of data recorded during the experiment. However the trial information data frame has one row for each trial.

Chapter 4

Studies in Predictive Eye-Hand Movement

In this chapter a series of studies focused on predictive eye-hand movements is presented. Each study has been designed to address questions related to the role of online and predictive mechanisms in visually guided control. In order to study different hypothesis concerning human visual-motor behavior, we created experiments in which subjects where immersed in a VR ball catching simulator. Use of virtual reality allowed us to record subjects ocular-motor behavior while they were free to move their head and hand in a natural setup. Thus increases the chances that our findings will generalize to the natural context. Here we briefly summarize the studies, and their contribution towards the understanding of predictive strategies for control of action.

In order to record subject's eye movements we used a built-in eye tracker inside head mounted display (HMD). However, our initial measurements revealed that the accuracy of the eye tracking system using only the calibration routine provided by the vendor is below the range of our scientific application. Therefore, the first study proposes a post-hoc calibration routine that reduces the spatial/temporal errors of a built-in eye tracking system throughout the data collection procedure.

The second study presents a VR ball catching study that the subjects were instructed to intercept a moving virtual ball while it disappeared midway through its parabolic trajectory. In order to successfully catch the ball, subjects had to adopt a predictive control strategy. We first investigate whether subjects engage in visual prediction when tracking the ball and if they did whether their predictive eye movements were correlated with their hand placement. We focused on subjects eye and hand movements when the virtual ball disappeared and show whether the temporal constraints of the task modulates the correlation between predictive eye and hand movements.

The third study utilizes machine learning models to test theories of visual prediction. Specifically, the model was designed to test the sufficiency of predictive control using a minimal representation between previously observed visual and kinesthetic information for explaining subject behavior made in prediction of a future state. This model is implemented as a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) architecture that uses current and previous information in order to predict the gaze-hand movement at a distant time into future. In addition, we investigate the significance of input source of information in guiding prediction.

Finally, in the fourth study, we investigate the influence of spatio-temporal constraints of ball trajectory on prediction. The reliability of visual information about ball kinematics varies during early compared to late portion of the trajectory. To investigate, we occluded the early or late portion of the ball trajectory and studied the change that it causes in eye-hand movement strategies compared to the no occlusion condition. Analysis of subjects eye-hand movement strategies brought insight to predictive strategies. Furthermore, we used subjects data to create an inverse reinforcement learning model of gaze-hand movements that captures both online and predictive components of the behavior. Characterizing subjects behavior using this model allows us to recover reward values that underlie visual and motor behavior during interception. The recovered rewards show the transition between online and predictive strategy in the course of a single trial. These results suggest that a single model of eye-hand movement is sufficient to capture both online and predictive component of behavior depending on the temporal constraints of the task and

whether the visual information is available or not. The following sub-sections present the studies conducted in order to understand human predictive strategies. Each of the following sub-sections present a conference proceeding or a published journal paper accordingly.

4.1 Study1: Eye Tracking Calibration in VR

Eye movement data is widely used to scientifically study human performance. The spatio-temporal accuracy of this data is intimately tied to the technology utilized for capture. Common eye tracking systems in use today are video-based pupil-center corneal-glint-reflection eye trackers. Such trackers can have varying degrees of intrinsic error depending on the properties of the hardware (cameras resolution, sampling frequency, etc.) as well as the data recording conditions (e.g., stimuli luminance and head movements) [29, 65]. Recent improvements in tracking technology integrate a 3D model of surface of the eye. Using multiple LEDs for illumination, this method can robustly estimate pupillary position based on the deformation of the corneal reflections. Depending on the research scenario, even small amounts of tracking error can have detrimental effects. For example, even a 0.5 offset of gaze location can lead to widely different conclusions about dwell time on areas of interest (AOI)[29].

Eye tracking systems have limited levels of accuracy that typically depend on a wide variety of physical, implementation, and environmental factors. For simplicity, many users rely on manufacturers' claims as to reliability and accuracy. However, the emergence of new technologies and curiosity about claimed versus actual, pragmatic accuracy has led to interrogation into the inherent properties of these systems [30, 66]. To minimize error, especially in scientific experiments, it is common to employ so-called characterization and/or calibration methods.

During a typical calibration, experimenters present the subject with some number of known-location visual stimuli. By comparing the gaze data with this 'ground truth' it is possible to construct some form of 2D transformation that minimizes the error between the observed and true locations[36, 67].Translational, rigid, similarity, affine, and linear fractional transformations are well suited for correcting common linear distortions introduced by physical and optical variability in eye tracking systems. Other methods, such as piece-wise transformations, topological / conformal techniques, statistical, and other model-based techniques may provide enhanced accuracy in

certain scenarios[68].

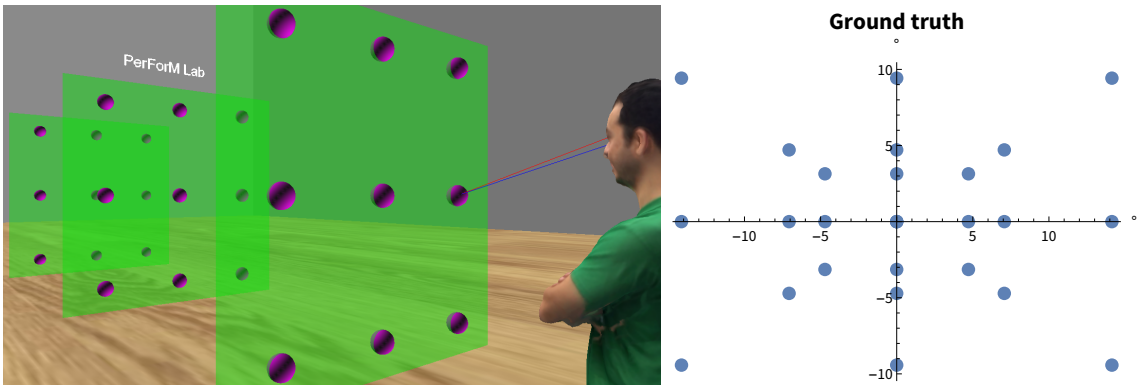
The focus of this study is to present a novel calibration method, suitable for Virtual Reality (VR) and other head-mounted eye tracking applications. Virtual reality techniques are increasing in popularity across many research disciplines[67, 69]. The addition of integrated eye tracking systems has provided a new perspective for vision scientists, allowing the study of the human visual system in a more naturalistic and less constrained environment. Tracking error is a crucial issue when wearing a Head Mounted Display (HMD) in a VR setting.Because subjects are free to move their head and explore while immersed in the virtual environment, a second source of error arises with the shifts and movements of the HMD. Our method uses binocular eye tracking of 27 spherical 'ground truth' objects at 3 different depths. We use well known 2D point matching algorithms to find a homography that minimizes the error at these different depths. We then extend this method from a static calibration to a dynamic one that takes place throughout the experiment.

Figure 4.1 illustrates the characterization setup. We refer to this grid of objects as the 'ground-truth', as viewed by a 3rd-person from inside the VR (left), and as projected onto the display of the HMD (right). Here the avatar represents the experimental subject, wearing an HMD, shown with binocular gaze vectors obtained from the in-HMD eye tracker. In the VR world the ground-truth occupies an equally spaced volume of $3 \times 3 \times 3$ spheres. This volume is located 2 m from the subject, spaced 0.6 m horizontally, 0.4 m vertically, and 2 m in depth. We scale the radius of the spheres such that they subtend an equal amount of visual angle ($15'$), regardless of depth or position. Their size in this illustration is exaggerated for the sake of clarity. These projected 2D points, as shown in the right-hand figure, are used as fiducial locations for point-matching transformations. The ground-truth grid is fixed with respect to the subject's cyclopean eye; such that the volume is always projected onto the same location on the HMD screen. The full experiment proceeds as follows. Before the experimental trials, 27 ground-truth locations are presented, one at a time for approximately 2s. For each location, we acquire 100 frames @ 60 Hz of binocular gaze data. We then perform the manufacturer's calibration procedure, and

subsequently repeat the 27 location presentation / collection task.

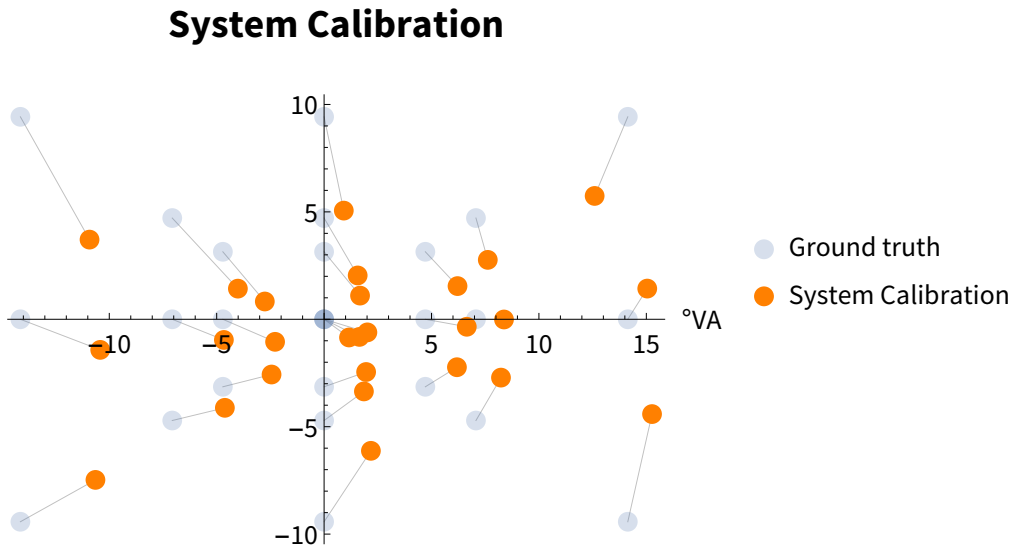
In order to properly characterize the system and acquire relevant experimental data we must calculate a 3D gaze vector into the VR world. Our calculation starts by placing a fiducial node into the 3D VR world coordinate system. We call this the 'cyclopean eye node'. Its location is continuously updated using motion capture data from markers placed on an Oculus HMD. The right and left eye nodes yoke to the cyclopean eye node, based on their inter-ocular distance (IOD).

Using ray-tracing and knowing HMD screen's distance to the cyclopean eye and its resolution, we can calculate the subject's gaze point on the screen. We refer to this as the point-of-regard or POR. Since the ground-truth objects are 'attached' to the subject's cyclopean eye, the true PORs remain the same regardless of head location or rotation. We calculate the gaze vectors for the right and left eyes using the aforementioned eye nodes, and calculate their respective PORs, along with cyclopean POR which is their average.



(a)

Figure 4.1: Calibration grid used for eye tracking error measurement



(a)

Figure 4.2: SMI calibration result

4.1.1 Spatial Characterization

The right hand side of Figure 4.1 shows the 2D POR data for all 27 ground-truth points. Figure 4.2 shows the results of the vendor-supplied system calibration with respect to these ground truth locations. Each dot-pair shows the reported location of the subject's gaze (orange) when looking at a particular reference ground truth location (purple). The central locations are reasonably well matched but the performance rapidly deteriorates as the location moves away from the center of the display. Furthermore, there is a slight vertical compression. The source of these distortions is somewhat unclear, but might be due to the system's optics and / or some form of non-linear screen distortion. The average angular error across the whole field of view is $\bar{x}_{err} = 2.9°, s_{err} = 1.6°$.

4.1.2 Point Matching and Transformations

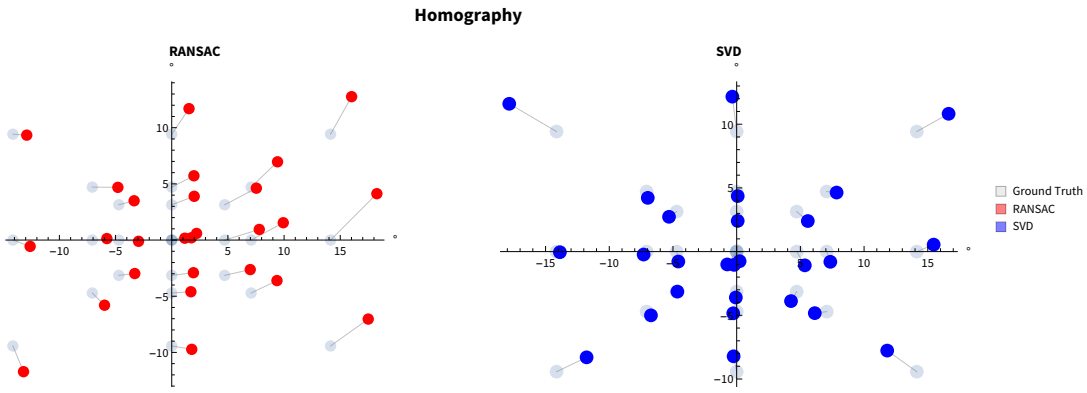
The problem consists of finding a transformation $\mathcal{T}$ such that locations in the observed locations $\mathcal{O}$ map as closely as possible to the ground-truth, $\mathcal{G}$. The result of this transformation is a form of isomorphism called a homography, $\mathcal{H}$ [68].

Random sample consensus (RANSAC) is an iterative method to estimate parameters of a mathematical model from a set of observed data which contains outliers [70]. In this case two sets of known ground truth and observed POR are the inputs and the output is a linear transformation that minimizes the error between two input sets. RANSAC is non-deterministic but able to deal with outliers and noisy data points efficiently.

Other methods, such as linear fitting of the locations via singular value decomposition (SVD) prove useful for scenarios such as ours where the distortions are largely linear and regular. In these algorithms, the absolute correspondence of points may or may not already be known. In our case, we know the corresponding data for each gaze position, making the task somewhat simpler. Here, we use RANSAC and SVD fitting to determine the transformation $\mathcal{T}$ to map $\mathcal{O}_{\mathcal{G}} \rightarrow \mathcal{G}$. We can then use $\mathcal{T}$ to create homographies $\mathcal{H}$ of other observed data $\mathcal{O}$.

4.1.3 Static Calibration Result

Figure 4.3 shows the result of our isomorphic corrections using the RANSAC and SVD methods described above. We compute these transformations on the already system-calibrated (e.g. 'uncorrected') locations. The nature of the two different methods is readily apparent - the RANSAC approach appears to attempt to broadly distribute the error across the entire space, while the SVD algorithm emphasizes the central locations over the peripheral. The resulting average angular error for the RANSAC method is $\bar{x}_{err} = 2.39^{\circ}, s_{err} = 1.11^{\circ}$, which is both lower and less variable than the vendor supplied results. The SVD method reduces the magnitude of error even further,



(a)

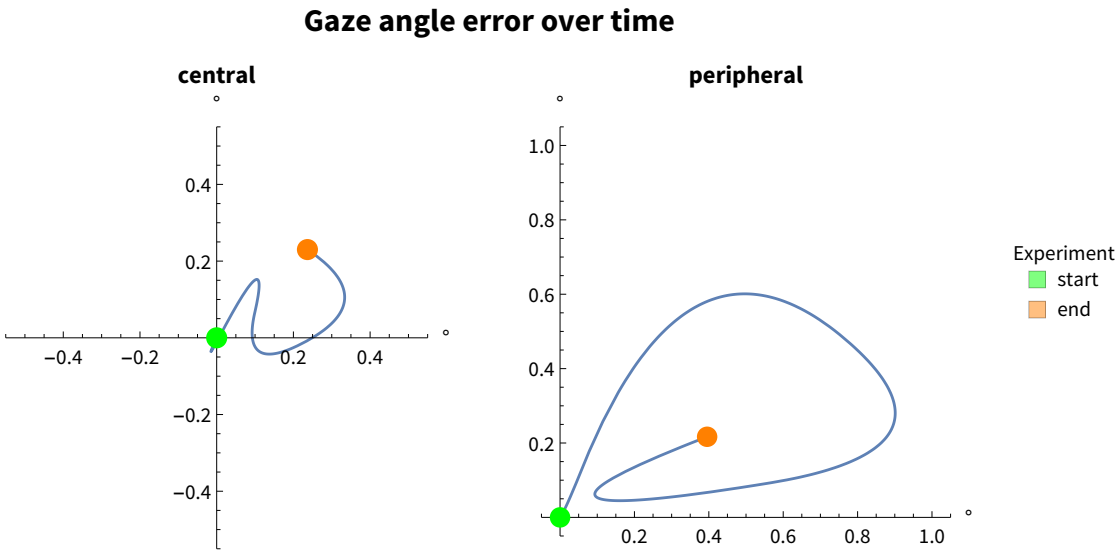
Figure 4.3: Homography Results

but with similar variance as RANSAC, $\bar{x}_{err} = 1.18°, s_{err} = 1.03°$.

Results demonstrate that we can improve on the vendor-supplied calibration by further transforming the resulting gaze data. Both the RANSAC and SVD homographies are the result of linear fractional transformations, computed using the vendor calibrated gaze positions. The RANSAC method results in an overall 18% reduction in error while the SVD method provides a further 51% reduction. Although these results promise to significantly reduce the angular error, a further challenge remains. Can we use this calibrated data throughout the timecourse of the experiment? Furthermore, can we compensate for variations introduced by the mechanical displacement of the VR HMD?

4.1.4 Dynamic Spatial Calibration

In a VR setup, depending on the nature of the experiment this could take up to 30+ minutes. The motion of the subject as well as potential calibration drift from the eye-tracker make our results less reliable as the experiment progresses. Here we propose a method for refining the



(a)

Figure 4.4: Gaze error for two calibration points in time

calibration over the entire timecourse of an experiment. Figure 4.4 demonstrates these two sources of error over time. To compute this, we showed 'ground-truth' locations at various timepoints throughout the experiment and asked the subject to fixate on them. We performed a time-varying cubic interpolation through the resulting gaze locations, resulting in the paths shown in the figure. The left graph shows the angular drift of the gaze when looking at a 'ground-truth' location at the center of the visual field. Whereas the right graph shows a point in the periphery. The drift of the central point wanders around the 'true' 0 location but the peripheral gaze wanders through almost a full degree.

One possible solution that is used previously is to perform 'recalibrations' during the experiment [71]. We use these recalibrations to 'reset' the isomorphic transform, thus minimizing error at punctuated times. This typically entails a delay in the task and can result in fatigue and loss of concentration. Here we perform a correction that, unlike our previous correction, changes contin-

Method	$\overline{x}_{err}$	s_{err}
System	2.90	1.60
RANSAC	2.39	1.11
SVD	1.18	1.03
Dynamic SVD	0.538	0.0668

Figure 4.5: Results for different calibration methods

uously throughout the experiment.

We use the same experimental task, apparatus, and virtual environment setup as shown in Figure 4.1.During the fixation period we acquire a batch of 100 uncorrected samples of the gaze at 60 Hz. We perform a Gaussian weighted average of these locations and classify this as the fixation location. As the experiment proceeds, a randomly selected ground-truth sphere appears in between batches of 3-5 non-calibration, ball-catching trials. We gather the same gaze information as in the initial presentation. This provides a time-varying set of uncorrected locations, relative to the ground-truth location. From this time series we can generate the paths of the uncorrected gaze, as shown in Figure 4.4, throughout the experiment.

By combining the uncorrected gaze time-series for each of the 27 ground-truth locations, we establish a function $\mathcal{P}(t)$ that returns the temporally interpolated grid of 27 gaze locations at time t. We then can use the point-matching and transformation methods described previously to calculate a unique homography, $\mathcal{H}_t$ for that specific time point. Thus, the corrections vary over time with the drift of the overall system. Based on the findings in the previous section, we chose a GLM / SVD method to compute the transformation, since this method reduced the overall angular error magnitude and variance the best.

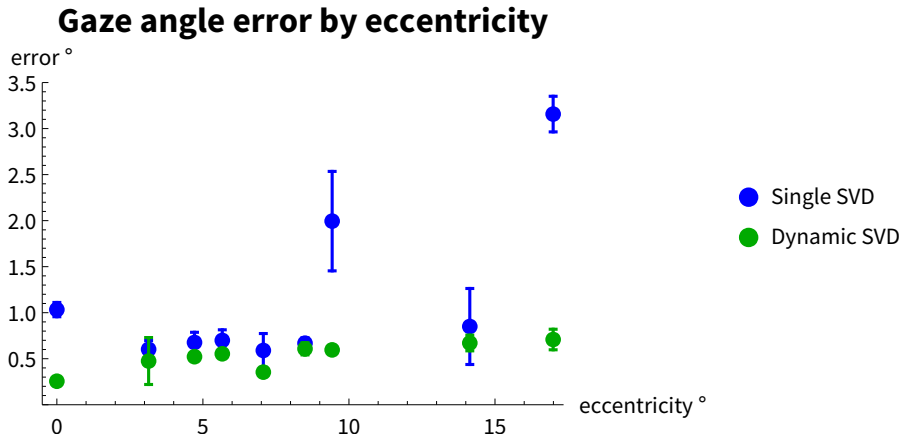
4.1.5 Dynamic Calibration Result

Table 4.5 shows the overall results for the dynamic technique, along with the previously described methods. Note that we average the error reported for the dynamic method over space and time. That is, the error at any particular t will be different by a small amount as the quality of fit of the transformation varies. However, as you can see, with an average magnitude of $\bar{x}_{err} = 0.538^{\circ}$ and a variability of only $s_{err} = 0.0668^{\circ}$ this method outperforms the static methods.

As we have previously noted, it is usually a good strategy to optimize the tracking accuracy in the center, e.g. 'straight ahead', viewing angle. The subject can easily turn their head in a VR environment to make a location 'more central' when necessary. Still, in a perfect world, the accuracy around the periphery should be high as well. Figure 4.6 shows a comparison of the static and dynamic methods' performance as a function of gaze angle eccentricity. The projection of the ground-truth spheres onto the HMD display occupied an region of roughly $\pm 16^{\circ}$ from the display center. For the best static correction (the SVD), shown in blue, error increases in both magnitude and variability as the screen locations become more peripheral. There is also slightly more central error. The dynamic correction reduces both the magnitude and the variability of the error over the whole range of eccentricities.

We set out to characterize the eye tracking error in our current VR HMD setup and to develop methods of minimizing these static and dynamic errors. Using transformations derived by comparing raw data with a ground truth we were able to reduce error significantly. By using the dynamic calibration method we were able to further reduce error that occurs as a result of movement and calibration drift over time by constructing a continuous mapping of transformations. These transformations are time-specific and compensate for instantaneous error as well as global, constant error.

For this experiment, we calculate the homographies off-line, after the experiment, for both the static and dynamic cases. It is possible to do some of this computation in real-time, especially with



(a)

Figure 4.6: Calibration Error vs Eccentricity

the assistance of GPU acceleration. However, since interpolation takes the 'future' locations of gaze position into account a different technique needs to be employed. We have experimented with Kalman-filter and 'hold-left' time-series approximations of our data and still are able to achieve satisfying results, though not quite as good as our off-line methods.

Interestingly, some of our exploratory results were improved when we used uncalibrated data from the tracker. This might be because in the proposed method, calibration points are presented at different depth, or vendor calibration might not be as compatible with the class of transformations (linear and partial linear) that are performed in this study.

4.2 Study2: A Common Predictive Strategy for Eye and Hand

4.2.1 Statement of the Problem

Although attempts to intercept a ball in flight are often preceded by predictive gaze behavior, the relationship between the predictive control of gaze and the effector is largely unexplored. The present study was designed to investigate the influence of the spatio-temporal demands of the task on a switch to the predictive control. Ten subjects immersed in a virtual environment attempted to intercept a ball that disappeared for 500 ms of its parabolic approach. The timing of the blank was varied through manipulation of the post-blank duration prior to the ball's arrival, and the shape of the trajectory was manipulated through variation of the pre-blank duration. Results reveal that the gaze movement trajectory during the blank was curvilinear, appropriately scaled to the curvature of the invisible moving ball, and the gaze vector was within 4 degrees of the ball upon reappearance, despite 10-13 degrees of ball movement. The timing of the blank did not influence the accuracy of predictive positioning of the paddle at the time of ball reappearance, indicated by the distance of the paddle relative to the ball's eventual passing location. However, analysis of trial-by-trial covariations revealed that, when the gaze vector more accurately predicted the ball's trajectory at reappearance, the paddle was also held closer to the ball's eventual passing location. This suggests that predictive strategies for paddle placement are more strongly mediated by the accuracy of gaze behavior than by the observed range of trajectories, or the timing of the blank.

Visually guided action involves both online and predictive components of control. For example, when attempting to catch a ball in flight, the control of hand position can be understood as an online coupling to visual sources of information available throughout the ball's trajectory [15]. However, investigations of the gaze behavior of individuals attempting to intercept a ball as it moves in depth suggest that this online component of control is accompanied by eye movements made in prediction of the ball's future trajectory [34, 59, 61–63, 72]. Under certain conditions, the predictive component can have a strong influence on factors leading to a successful intercep-

tion of a target moving in two dimensions [73], perhaps due to extra-retinal contributions from smooth pursuit before visual feedback about the moving target's position is removed (e.g. through occlusion, or target blanking).

This study has been designed to investigate the factors that mediate the strength of the relationship between visual prediction and movements of the hand and body when attempting to intercept an object moving in depth. Although prediction may be accomplished through a "prospective" coupling of behavior to visual information that forecasts a likely future state [74], this study more specifically focuses on predictive behavior that is separated in time from the sensory information that informed the control strategy. Our approach to studying prediction was to immerse subjects in a virtual reality (VR) environment in which the task was to use a hand held, motion-tracked badminton paddle to intercept a launched ball moving along a parabolic flight over a distance of approximately 20 m to a location within the subject's reach (Figure 4.7). Upon successful interception, the virtual ball would stick to the paddle for a brief duration before the end of the trial, providing visual feedback about the accuracy of paddle placement before the ball would disappear, and the next trial would begin. To promote predictive control strategies for gaze and the hand/paddle, the ball was made invisible (or *"blanked"*) for 500 ms on each trial during its parabolic flight towards the subject. The duration between the ball's reappearance and its arrival at the subject or, post-blank duration was limited to 500, 400, or 300 ms.

Our first hypothesis is that participants will engage in visual prediction through the blank period. This assumption is supported by several studies of gaze behavior prior to interception in naturalistic conditions [34, 61, 62, 75, 76]. The accuracy of the visual prediction will compare the position and velocity of the subject's gaze behavior at the end of the 500 ms blank to the position and velocity of the ball at the same frame: at the time of reappearance, prior to the availability of post-blank visual information, and shortly before the attempted interception.

Our second hypothesis is that the participant's strategy for positioning the paddle during interception will transition from an online mode of control to a predictive mode of control when the

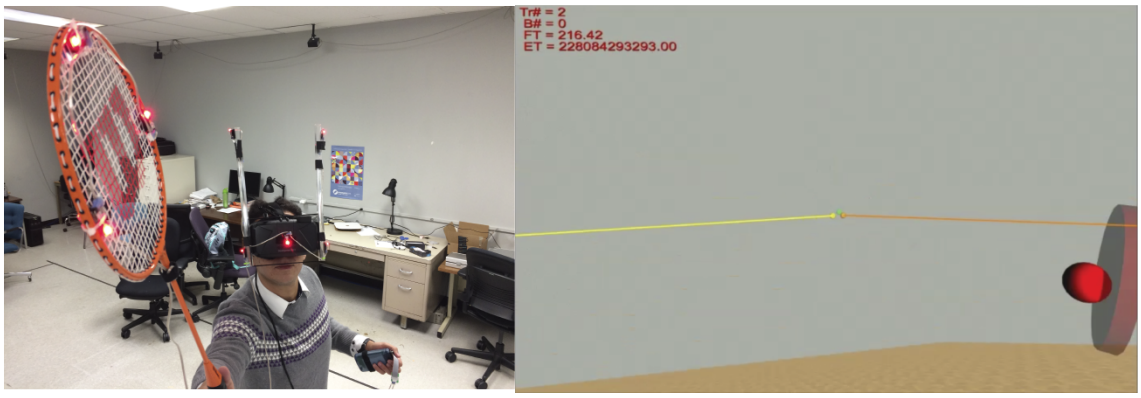


Figure 4.7: **A.** Subjects wore an Oculus DK2 with integrated SMI eyetracker with a sampling rate of 60 Hz. **B.** The experimenter's desktop view of what the subject saw inside the helmet. The lines receding in depth represent the left and right gaze vectors, and the text in the upper left encodes trial number, block number, and time-stamps. The red disc represents the face of the paddle, on which can also be seen the red virtual ball.

post-blank duration is shortest, and there is very little time for online control after the ball reappears. This hypothesis is motivated, in part, by the finding that, when visual feedback is removed upon the initiation of a movement to intercept, subjects' transition between online and closed loop control strategies is modulated by the quality of prediction [77, 78]. In the context of natural(istic) interception tasks, the quality of prediction degrades quickly when the target is occluded for the final 280 ms of its flight or longer [77, 79]. Within the context of the present study, a prediction of the post-blank ball trajectory would have to be made on the basis of pre-blank visual information that is at least 500 ms old, and thus performance based on a prediction through the blank period is expected to be degraded relative to natural, online control. It follows that it would be advantageous to reestablish an online control strategy after the blank if the post-blank duration is sufficiently long for motor planning and execution on the basis of post-blank visual information. If this is not possible, subjects may adopt a more predictive mode of control on the basis of pre-blank visual information. This would be evident if more dramatic paddle movements occur during the blank, and if the relative distance of the paddle from the ball's eventual passing location at the end

of the blank period is lower when the blank occurs later in the trial.

Our third hypothesis is that visual and motor strategies are driven by shared resources. If so, we predict that predictive movements of gaze and the paddle will covary on a trial-by-trial basis. Similar signs of visuo-motor coordination were apparent in a real-world interception task involving a carefully calibrated ball launching device [63]. Although the duration of tracking covaried with interception performance on a trial-by-trial basis, the effects were lost when averaging within conditions. Although the authors speculated that the influence of visual prediction on the subsequent movement to intercept is contingent upon the spatio-temporal demands of the task, the experiment was not designed to specifically test this hypothesis. In the present experiment, the spatio-temporal task demands are modulated by varying the timing of the blank relative to the ball's arrival, which is accomplished through manipulation of post-blank duration. Therefore, if visual and motor strategies involve shared resources, we would expect errors to be coupled at low values of post-blank duration, when the task places high temporal demands upon the participant, and elicits a more predictive mode of control for both gaze and the paddle.

It is notable that, because the blank duration is fixed, reductions in post-blank duration will also have the effect of reducing the ball's overall time-of-flight, and bring about a straighter (less curved) ball trajectory. As a result, any conclusions concerning the role of visual prediction would be confounded by qualitative changes in the ball trajectory. To test the contribution from changes in the ball's trajectory to visual prediction, the duration between launch of the ball and the onset of the blank period, or pre-blank duration, varied between three-values, between trials. These changes to the pre-blank duration modify the characteristics of the ball's trajectory independently of post-blank duration. Combinations of trajectories are presented in Figure 4.8.

In summary, this experiment has been designed to investigate the role of visual prediction in guiding the subsequent paddle movement in a VR interception task in which the ball is "blanked" for 500 ms as it moves in depth. We have put forth the following hypotheses: 1) that subjects will engage in visual prediction through the blank period, 2) that subjects will engage in predic-

tive movements of the paddle only when the spatio-temporal demands of the task prevent online control, 3) that when subjects are engaged in both visual and motor control, their movements will demonstrate correlated errors between the modalities.

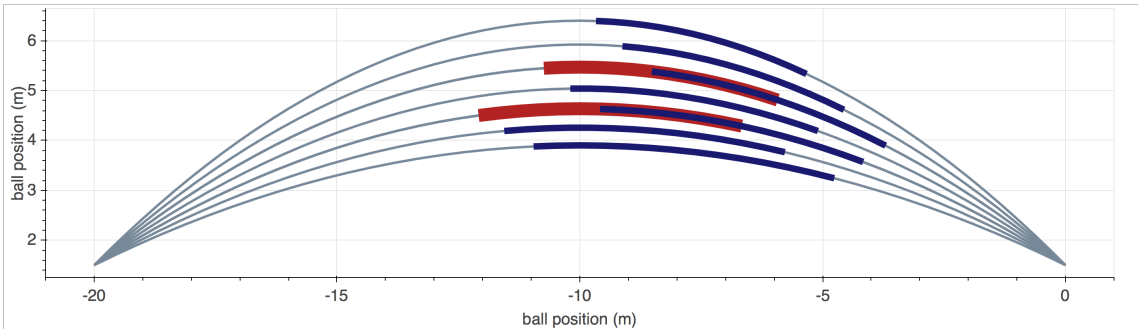


Figure 4.8: A side-view of the trajectories used in the experiment, which were comprised of a blank period of 500 ms (thick regions), a pre-blank duration of 600, 800, or 1000 ms, and a post-blank duration of 300, 400, or 500 ms before the ball's passage over the X axis upon which the subject was standing. Thick regions along the trajectories represent the timing of the blank period. Overlapping blank periods reflect the fact that some trajectories have a common overall time-of-flight despite differences in pre-blank duration and TOR, resulting in two-possible times at which the blank may occur. Although ball trajectories in the experiment could have approached from a variety of angles, this representation assumes a single approach angle for visual simplicity.

4.2.2 Methods

The ten participants (7 male, 3 female) were between 19-30 years of age and had normal vision in the absence of visual correction. This study was approved by the Institutional Review Board at the Rochester Institute of Technology.

Experimental Apparatus

Stimuli were delivered by an Intel i7-based PC with an NVIDIA GTX 690 connected to the Oculus Rift DK2 head mounted display at 75 Hz, and an NVIDIA GTX 760 connected to the experimenter's desktop display. The computer ran Windows 7, and the virtual environment was rendered

using the Vizard Virtual Reality toolkit by Worldviz. Physics were simulated using the OpenODE physics engine so that ball trajectories matched those expected within a real-world environment in the absence of wind resistance (Figure 4.7).Collisions between the ball and the paddle or other surfaces was also detected using the physics engine. To improve the fidelity of the physical simulation and collision detection, the physics engine was updated 10 times between frames (i.e. 750 Hz).

The Oculus Rift HMD has an approximate field of view of $100°$, and an approximate angular resolution of 10-15 pixels per degree (dependent upon the position of the eye inside the hmd). Head and paddle position/orientation were sampled and recorded at 75 Hz using a 14 camera *Phasespace* X2 motion capture system, with a measured latency between the time a sensed movement would be reflected on the display of less than 30 ms. Eye movements were recorded with an *SMI* binocular eye tracker sampled at 75 Hz. A post-hoc correction was applied, as described in [48], to correct for helmet slippage and other sources of spatial error in the eye tracking signal. The average eye tracking accuracy after calibration and correction was $0.53°$ for the central visual field ($FOV < 10°$) and $2.51°$ in the periphery ($10° < FOV < 30°$).

Experimental Design

Participants were instructed to perform the interception task using their dominant hand, and all participants self-selected to hold the paddle with their right hand. All subjects conformed to the verbal instructions to catch, and not to hit the ball, using the paddle. Upon collision, the ball would stick to the virtual paddle for a brief duration before the end of the trial, providing visual feedback about the accuracy of paddle placement before the ball would disappear, and the next trial would begin.

The red virtual ball was launched from a plane that was 6 m wide $\times$ 1.5 m high and parallel to both the virtual room's X-axis and the vertical axis. The launched ball's trajectory was calculated so that, if the ball were not intercepted, it would pass through a location randomly selected from a 1

m $\times$ 1 m plane near the subject, also parallel to the room's X-axis and the vertical axis. Trajectories consisted of a pre-blank duration of flight (600, 800, or 1000 ms), a 500 ms blank duration, and a post-blank duration of 300, 400, and 500 ms before the time at which the unimpeded ball would pass the X axis at which the subject was standing. This design produced 9 combinations of pre and post-blank duration, and 7 possible flight-durations (Figure 4.8). Each subject performed 135 ball catching trials.

Data preparation and analysis

Data preparation and non-statistical analysis of gaze and paddle movement data was conducted with Python 2.7 and 3.4, with modules Numpy, Pandas for computation, and a mixture of Bokeh and Plotly for figure generation.

Analysis began by averaging the left/right eye-in-head vectors provided by the SMI eye tracker to produce a single unit vector that represents the cyclopean gaze direction. The transformation matrix of the motion-tracked head was applied to the vector to cast its appropriate location in front of the head in the virtual world. Data was passed through a median filter (length 3). The smoothed velocity signal was calculated with a third-order Savitsky-Golay filter with a window size of 5.

Catching Error: Catching error was calculated as the distance of the ball from the center of the paddle at the moment that the ball either collided with the paddle, or the moment that the ball passed by the vertically oriented plane located at the face of the paddle and orthogonal to the ball's trajectory.

The influence of the timing of the blank period on catching behavior: Special steps were taken during the design phase to facilitate an investigation of whether the timing of the blank period within a particular trajectory affected behavior. This was made possible because both flight distance and blank duration are constant across all trial types. As a result, the shape of the ball's trajectory is determined by overall time-of-flight, which is equal to the sum of the pre-blank dura-

tion, blank duration, and post-blank duration. The design was such that there were two ”paired” conditions that were identical in time-of-flight (and thus ball trajectories), but that differed in the timing of the blank period. These two pairs are both indicated in Figure 4.8 by the third and fifth lines from the top; for each pair, the partially overlapping bold portions of the trajectory indicate the two possible timings of the blank period within the otherwise identical trajectories. The first pair shared an overall time of flight of 1800 ms, and included 1) the condition with a pre-blank duration of 1000 ms and a post-blank duration of 300 ms, and 2) the condition with a pre-blank duration of 800 ms, and a post-blank duration of 500 ms. The second pair shared an overall time of flight of 1600 ms, and included 1) the condition with a pre-blank duration of 800 ms and a post-blank duration of 300 ms, and 2) the condition with a pre-blank duration of 600 ms, and a post-blank duration of 500 ms.

Visual prediction error: To measure the accuracy of the subject's prediction of the ball's displacement at the time of reappearance, we measured the angular distance between the gaze vector and the unit vector extending from the eye to the ball (i.e. the eye-to-ball vector) at the last frame of the blank period.

Pursuit gain: If subjects adopted the strategy of engaging in visual pursuit/tracking of the ball across the blank period and leading up to an attempted interception, the speed of rotation of the gaze vector around the cylcopean eye would be well matched to the speed of rotation of the vector extending from the cyclopean eye to the ball around the cyclopean eye at frame of ball reappearance. Pursuit gain represents the ratio of these two values, where a ratio less than one would indicate that the gaze vector was rotating more slowly than the ball. The angular velocities of the gaze-in-world and eye-to-ball vectors were measured by calculating the vector distance between subsequent samples of the filtered signal, and then dividing the duration per frame, yielding a measure of velocity in degrees per second.

Polynomial complexity of the gaze and ball's trajectory during the blank: Predictive strategies proposed in the literature span a continuum from those that rely upon complex internal models of object dynamics, to simplistic heuristic strategies that operate in the absence of stored priors [15]. In the present task, the complexity of the predictive gaze behavior demonstrated during the blank can provide evidence that may help disambiguate between these strategies for prediction. For example, if gaze trajectories are linear even in the presence of non-linear ball trajectories, it would suggest that either a simple heuristic strategy is sufficient to guide behavior given the constraints of the task. In contrast, a complex, non-linear gaze movement that is well matched to that of the (invisible) ball would suggest the influence of a learned model of target movement.

To assess the complexity of predictive gaze behavior during the blank, first-order through fourth-order polynomial models were fit to the trajectory of the gaze vector through spherical space during the blank. Models were constrained so that the intercept aligned with the azimuth and elevation of the initial gaze sample at the time that the ball disappeared. To assess whether increasing the polynomial order was justified, the winning model was selected on the basis of adjusted R^2, a measure of goodness of fit that avoids over-fitting by incorporating a penalty for additional model terms. To assess whether the curvature of the gaze vector was well suited to that of the ball, the same procedure was applied to the ball's trajectory through head-centered spherical space, and the polynomial orders of the best fitting models for gaze and the ball were compared on a trial by trial basis.

Paddle velocity ratio: A measure of relative paddle velocity during and after the blank is used assess whether subjects guided paddle positioning in an online manner on the basis of post-blank visual information, or whether the paddle is positioned in a predictive manner, during the blank period. The mean paddle velocities for the two periods are calculated and then expressed as a ratio. The post-blank period is defined as the region from when the ball reappeared, until the earlier of

two events: the moment that the ball either collided with the paddle, or the moment that the ball passed by the vertically oriented plane located at the face of the paddle and orthogonal to the ball's trajectory.

Predictive catching error: The measure of predictive catching error is used to assess the accuracy of the positioning of the paddle at the moment of ball reappearance relative to the ball's post-blank trajectory. Paddle position is sampled at the time of ball reappearance. The ball's trajectory is then extrapolated until it collides with the vertical plane that passes through the center of the paddle and that is orthogonal to the ball's trajectory. The predictive catching error is the two-dimensional distance along this plane from from the center of the ball to the center of the paddle, measured in meters.

Statistics

Univariate ANOVA were conducted using JASP version 0.9.1, and violations of sphericity were subjected to Greenhouse-Geisser correction. Prior to analysis, normality of the residuals was assessed through use of the Shapiro-Wilk test, and no violations were found. Significance was assessed using an alpha-level of 0.05 with Bonferroni correction for family-wise error rates.

Mixed linear models were used to assess the covariation of the continuous variables of predictive catching error (pce) and visual prediction error. The mixed linear model is presented in equation 4.1:

$$\sqrt{pie_{si}} = \beta_0 + \beta_1 \text{vpe}_{si} + \beta_2 \text{pre}_{si} + \beta_3 \text{post}_{si} + \beta_4 \text{vpe}_{si} \times \text{post}_{si} + S_{0s} + \epsilon_{si} \qquad (4.1)$$

Heteroskedasticity and normality were assessed through visual inspection of the residuals [80], and these violations were mitigated via box-cox analysis [81] and square-root transformation of the latent response variable, predictive interception error (pie). Model terms represent the categorical

fixed effects of pre-blank and post-blank duration (pre and post), the continuous predictor visual prediction error (vpe), and the interaction of post-blank duration with visual prediction error. All values varied with item i and subject s. Mixed model analysis was conducted in R Studio version 1.136. The model was implemented using function lmer of the package LME4. Significance tests for fixed effects were run using the package lmerTest, and the related figures were generated using the library ggplot.

An identical model was used to evaluate the relationship between predictive catching error and the pursuit gain (pg), and this model is presented in Equation 4.2. In response to similar signs of heteroskedasticity, the response variable received the same square-root transformation prior to model fitting.

$$\sqrt{pce_{si}} = \beta_0 + \beta_1 \text{pg}_{si} + \beta_2 \text{pre}_{si} + \beta_3 \text{post}_{si} + \beta_4 \text{pg}_{si} \times \text{post}_{si} + S_{0s} + \epsilon_{si} \tag{4.2}$$

The contribution of each fixed effect was evaluated through a comparison between models of increasing complexity. For example, the first test of the model presented in equation 4.1 involved a comparison of a model which included only the random effect of subject to a model that also included the fixed effect of visual prediction error. The significance of the fixed effect was evaluated using a Chi-Squared likelihood ratio test. In addition, we required that, for a model to be accepted as superior, it must also reduce the Akaike Information Criterion (AIC) - a widely accepted metric used to assess whether the increase in model fit is justified by the inclusion of the additional terms. If the term passed these criterion its effect was considered significant, and the term was incorporated into the final model.

4.2.3 Results and Discussion

Interception rates

Interception was more difficult at shorter post-blank durations.

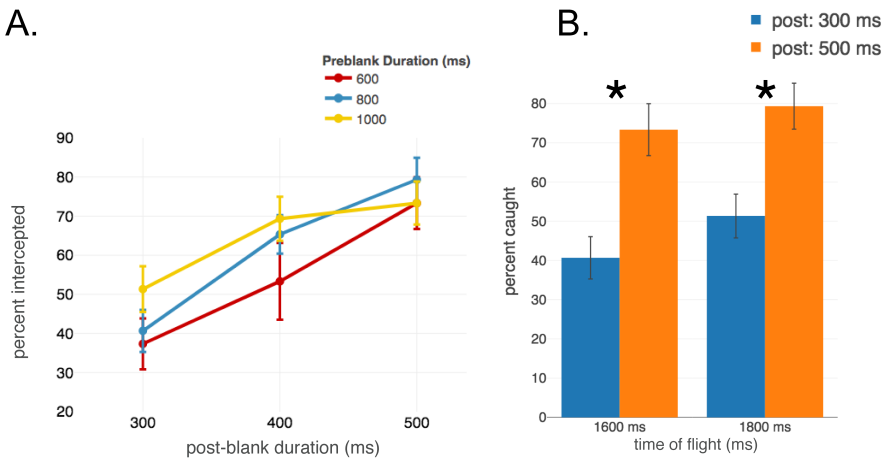


Figure 4.9: **A.** Interception rate by pre and post-blank duration. Error bars represent 95% confidence intervals reflecting within subject variability. **B.** A comparison of interception rate for the two pairs of conditions within which the trajectory is constant, but timing of the blank differs (see also Fig 4.8).

Interception rates are represented in Figure 4.9. Rates were lowest at shorter post-blank durations, when the ball appeared latest. A comparison between "paired" conditions that were identical in trajectory (as indicated by time-of-flight), but that differed in the timing of the blank period, reveals that subjects had a more difficult time intercepting the ball when it reappeared later. Figure 4.9B. presents two of these pairs. Within the pair that shared a time-of-flight of 1800 ms, the condition in which the ball reappeared later (pre: 1000 ms, Post: 300 ms) had an interception rate near 50%. In contrast, the condition with an earlier blank period (pre: 800 ms, post: 500 ms) had an interception rate near 80% ($t(9.00)=3.43$, $p<0.004$, $r=0.75$). Within the pair that shared a time-of-flight of 1600 ms, subjects caught approximately 40% of balls with a later blank period (pre:

800 ms, post: 300 ms) and above 70% for balls with an earlier blank period (pre: 600 ms, post: 500 ms; $t(9.00)=2.93$, $p<0.01$, $r=0.7$). Overall, this suggests that subjects had a more difficult time intercepting the ball when the blank period occurred closer to its arrival, and thus the ball reappeared closer to the participant's head (Figure 4.9A). A comparison between identical trajectories that differ only in the timing of the blank suggested that this effect could not be attributed to the shape of the trajectory or the ball's time of flight (Figure 4.9B).

Gaze behavior

Figure 4.10 presents trajectories of the ball and gaze through head-centered spherical space in which the azumithal plane passed through the subject's eye and parallel to the ground plane, and the elevation plane passed through the gaze vector the world's vertical axis. This figure is meant to be representative of the ball trajectories and typical subject behavior. During and shortly following the blank period, gaze is characterized by a combination of smooth pursuit and catch-up saccades. Visual tracking behavior ends prior to the ball's arrival, during the ball's high-velocity movement through spherical space.

Subjects compensated for the ball's angular displacement during the blank

Figure 4.11 presents the angular distance between the gaze vector and the ball's position throughout the last portion of the ball's trajectory, broken down by the post-blank duration. Regardless of condition, mean error remained below 4 degrees throughout the blank period. Analysis of the ball's displacement (Figure 4.12A) reveals that this error reflects partial compensation for movement of the ball during the blank. The magnitude of angular ball displacement during the blank period ranged from 10.4-12.6 degrees. Figure 4.12B presents the angular distance between the gaze vector and the edge of the ball at time of its reappearance (i.e. visual prediction error). This measure suggests that movement of the gaze vector during the blank resulted in a mean end-point

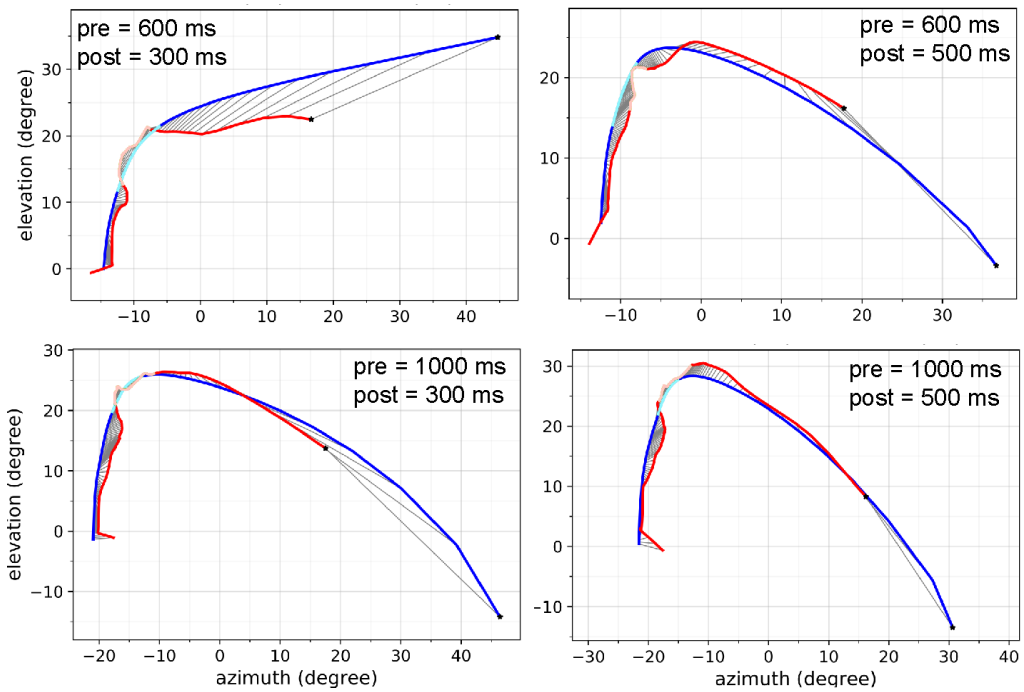


Figure 4.10: The trajectory of the ball and gaze vector through head-centered spherical space. These representative trials are from one subject for each combinations of pre-blank duration of 600 or 1000 ms, and post-blank duration of 300 or 500 ms. The visible portion of the ball's trajectory is represented in blue, and the portion of the ball's trajectory for which it is invisible is in cyan. Gaze is represented by the red trajectory, and the portion of the gaze trajectory when the ball is invisible is in pink. Grey lines connect samples of gaze/ball trajectory taken on the same frame. Trajectories move from left to right and end at the moment the ball either hits the paddle, or passes by a plane at the paddle's face.

accuracy of approximately 2.5-4 degrees upon the ball's reappearance (Figure 4.12B). Neither main effects of pre-blank duration, post-blank duration, nor the interaction reached statistical significance (pre: $F(2,18)=1.17$, $p=0.33$, $\eta_p^2=0.12$, post: $F(2,18)=0.64$, $p=0.64$, $\eta_p^2=0.05$, Interaction: $F(2.12,19.03)=1.02$, $p=0.41$, $\eta_p^2=0.10$).

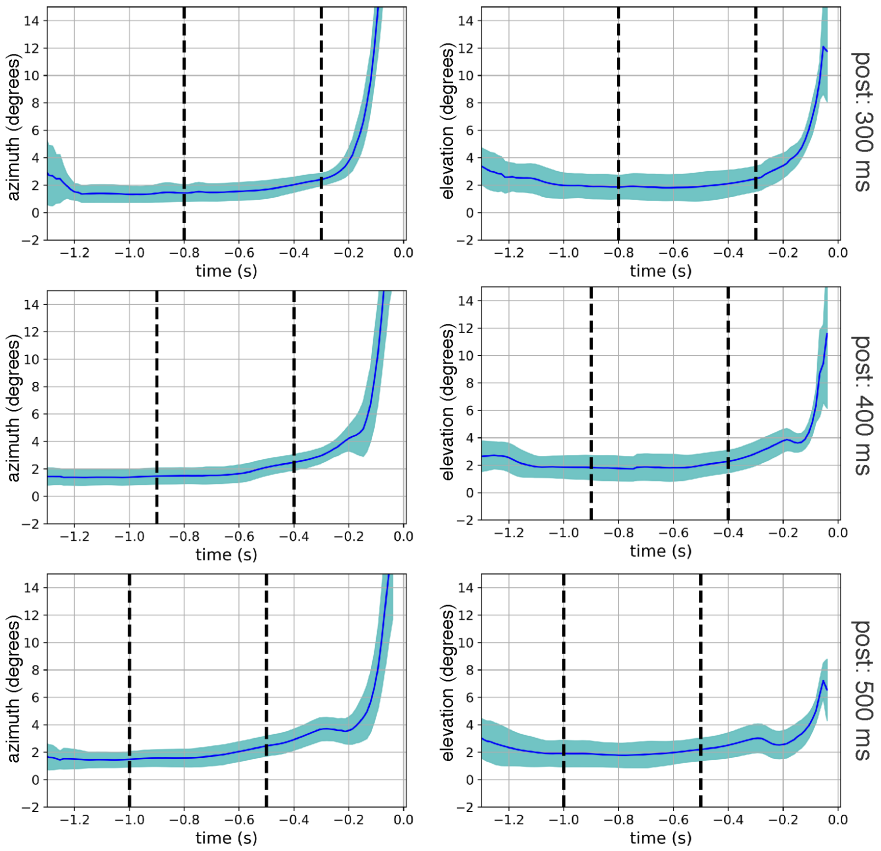


Figure 4.11: Mean angular distance between the gaze vector and the ball over the course of each trial type. Trials within each condition are aligned by the start and end of the blank duration, as indicated by vertical dashed lines. Shaded regions indicate 95% confidence intervals with between-subjects error removed.

Pursuit gain dropped below 1 for balls that were moving quickly upon reappearance.

Angular velocity of the ball around the head is presented in Figure 4.13A. This figure shows that, at lower values of post-blank duration, the ball was moving at higher angular velocities. This is reasonable when one considers that, even if a ball that is moving at a fixed velocity through euclidean space, if it reappears later and closer to the subject's head it will be moving more quickly

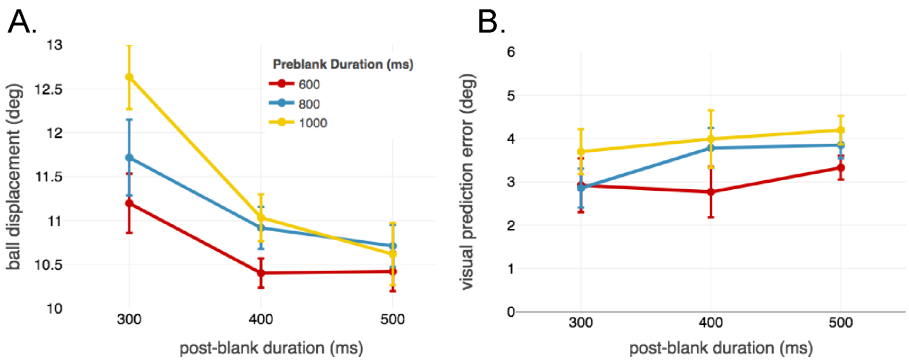


Figure 4.12: **A.** Angular displacement of the ball as it moved around the subject's head during the 500 ms blank period for which the ball was invisible. **B.** Angular distance between the gaze vector and the ball at the time of ball reappearance. Error bars indicate 95% confidence intervals with between-subjects deviation removed.

through head-centered spherical space (measured in degrees azimuth/elevation). Consequently, and as is shown in Figure 4.13B, subjects had a more difficult time pursuing the ball at lower values of post-blank duration. In contrast, subjects were generally able to scale pursuit velocity to the ball's velocity at higher values of post-blank duration, when the ball reappeared earlier and farther away from the subject's head ($F(1.25,11.27)=19.86$, $p<0.001$, $\eta_p^2=0.69$). Following Bonferroni correction for family-wise error, neither the main effect of pre-blank duration nor the interaction reached significance ($F(2,18)=4.23$, $p=0.031$, $\eta_p^2=0.32$); $F(4,36)=1.35$, $p=0.24$, $\eta_p^2=0.15$). Post-hoc tests revealed differences between 300ms-400ms ($p<0.011$), 300-500 ms ($p<0.003$), and 400-500 ms ($p<0.01$).

The curvature of gaze during the blank was tailored to the stimulus.

Polynomial models of fit to the trajectory of the gaze vector during the blank reveal that prediction was not a simple linear extrapolation of the ball's movement before the blank. Figure 4.14 presents an example of two trials in which models varying in their polynomial complexity were fit to the gaze trajectory. For each trial, the "winning" model was selected on the basis of adjusted R^2, a

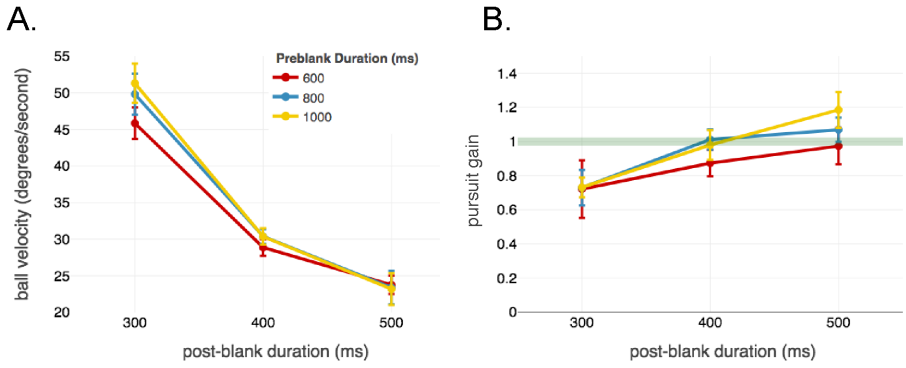


Figure 4.13: **A.** Angular velocity of the ball around the head upon reappearance. **B.** The ratio of angular velocity of the gaze vector and the eye-to-ball vector upon ball reappearance, or pursuit gain. Values reflect the average ratio within a 100 ms window centered upon the time of ball reappearance. The horizontal line indicates unity gain. Error bars represent 95% confidence intervals, with between-subject variability removed.

measure of goodness of fit that penalizes for the incorporation of additional terms, and overfitting. For example, because the fourth order polynomial model presented in the top row of Figure 4.14A provides a better overall fit to the gaze data, one would expect a higher R^2, however, the adjusted R^2 is slightly lower than the second-order polynomial model.

To assess whether the curvature of the the gaze trajectory was appropriately matched to the curvature of the ball, the same procedure was subsequently applied to the ball's trajectory through spherical space, and trials were binned in a two-dimensional histogram according to the complexity of the polynomial models needed to account for both gaze and ball trajectory (Figure 4.15). For example, consider that for a trial in which the pre-blank duration was 600 ms, and the post-blank duration was 300 ms, the trajectory of the ball was best approximated by a first order polynomial model, and the trajectory of gaze by a fourth order polynomial. This trial would contribute one count to the top-right bin of the upper left subplot. Brightness reflects the cumulative sum of all counts within a single bin, normalized within each combination of pre and post blank duration.

The concentration of data points along the second-order bin on the vertical (ball) axis in all

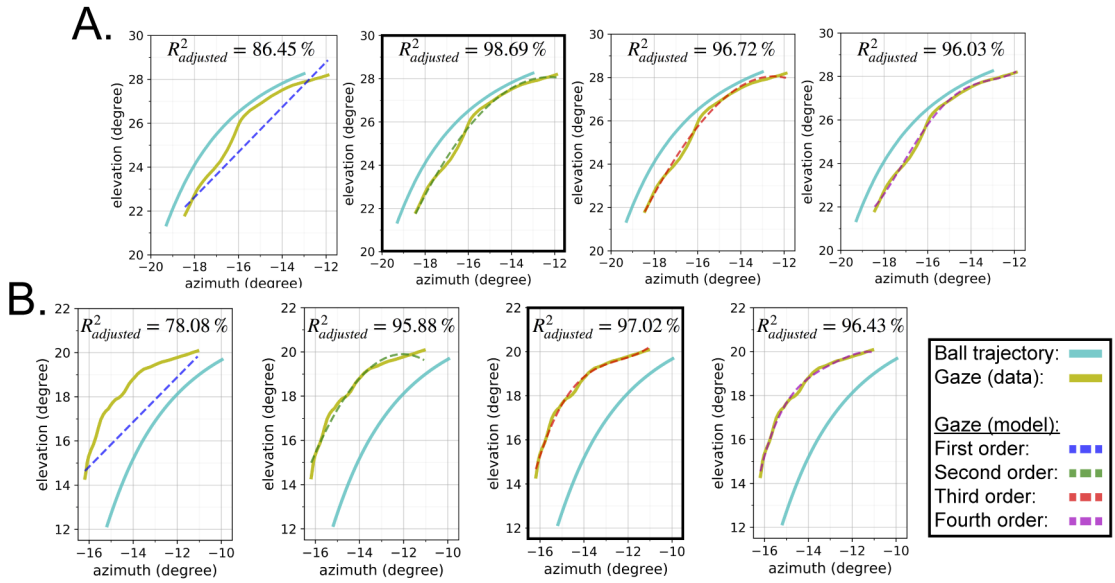


Figure 4.14: The trajectory of the ball and gaze through spherical space for two representative trials. Several polynomial fits that vary in complexity are overlaid upon the gaze trajectory. **A.** A trial in which gaze was best approximated by a second order polynomial. **B.** A trial in which gaze was best approximated by a third order polynomial.

subplots of Figure 4.15 reveals that the ball's trajectory through spherical space is best fit by a second-order polynomial in all conditions. Similarly, the curvature of the gaze trajectory is best explained by a second-order polynomial model in all conditions. Together, this reveals that that the gaze trajectory during the blank period was non-linear and appropriately complex given the non-linear trajectory of the ball as it moved through spherical space.

Summary of gaze behavior

Analysis of the stimulus and of gaze behavior as balls moved through head-centered, spherical space revealed that subjects regularly engaged in visual prediction during the blank period. Prior to and during the blank period, the subject maintained less than four degrees of distance between the gaze vector and the vector extending from the eye to the ball. During the blank period, this low-

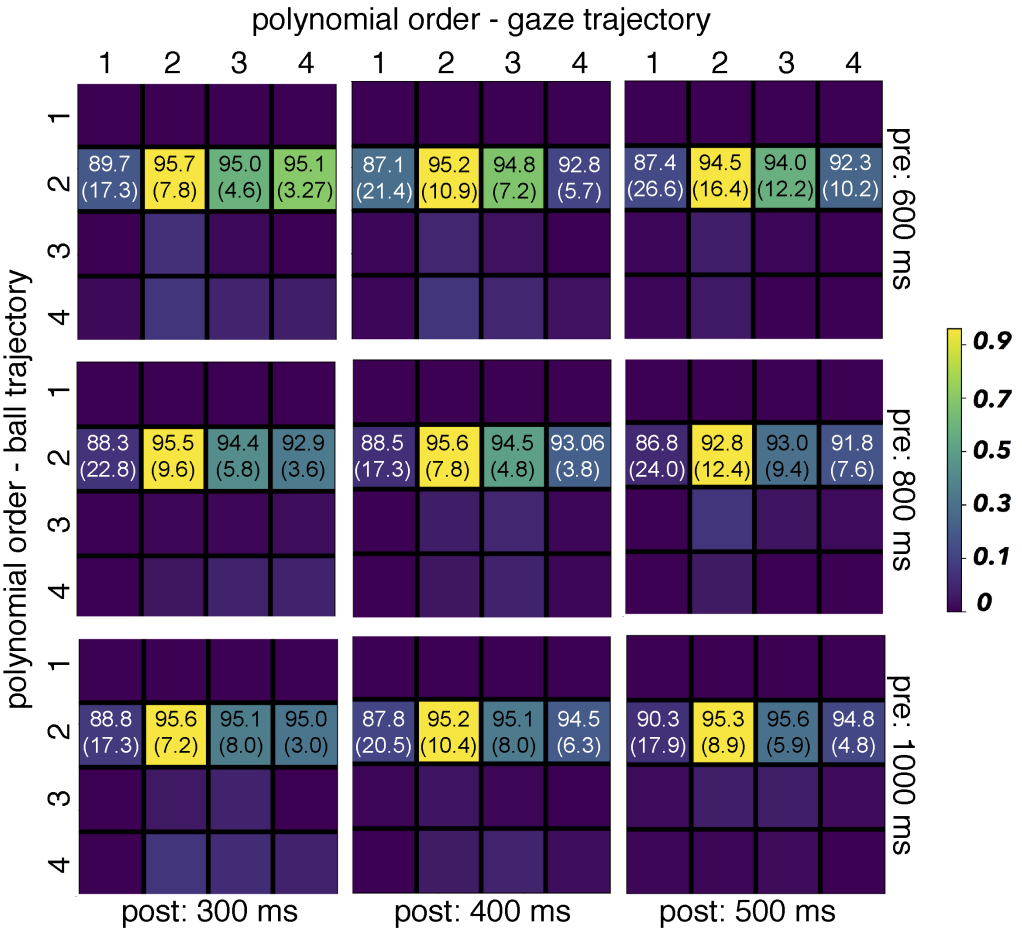


Figure 4.15: A 2D histogram which represents the complexity of the polynomial models needed to model the trajectory of the gaze vector and the ball through spherical space (as in Figure 4.14). Cell brightness reflects the probability of each bin's associated polynomial pair, normalized within each combination of pre and post-blank duration. In addition, numerical insets show the median and 95% confidence interval of adjusted R^2 for a subset of bins.

error was made possible by accurately accounting for the curvature of the ball's path on the basis of pre-blank visual information. Although positional accuracy was maintained across conditions, pursuit gain dropped from values near to 1 to values near to 0.7 when the ball reappeared later in the trial, as the result of a later blank period. This likely reflects increased task demands; when the blank period occurs later in the trial, when the ball is closer to the head, it will undergo a greater displacement during the blank, and be moving at a greater velocity upon reappearance.

Paddle placement and movement kinematics.

The speed of the paddle relative to its arrival for a subset of conditions is represented in Figure 4.16. Visual inspection of these velocity profiles suggests that the speed of paddle movement was greater after the blank when the blank occured earlier relative to the ball's arrival (i.e. at longer post-blank durations; right column of Figure 4.16). In addition, paddle speed was greatest during the blank for the condition with the shortest time-of-flight, when the ball disappeared later in its trajectory (pre=600 ms, post=300ms, top-left in Figure 4.16). These figures indicate that the subject chose to move the paddle during the blank, possibly on the basis of a prediction formed from pre-blank visual information.

To investigate whether mid-blank paddle movements were more dramatic at low values of post-blank duration, when conditions expected to elicit more predictive behavior, we investigated the average ratio of paddle velocity during the blank and post-blank period (Figure 4.17A). In all situations except the condition with the shortest time-of-flight (pre: 600ms, post:300ms), the mean velocity was higher during the post-blank period than during the blank period (i.e. the ratio was <1). Although the main effect of pre-blank duration was not significant ($F(1.3,11.7)=4.10$, $p<0.058$, $\eta_p^2=0.728$), the main effect of post-blank duration was significant, ($F(2,18)=24.115$, $p<0.001$, $\eta_p^2=0.728$), as was the interaction of pre and post-blank duration ($F(1.8,16.5)=8.59$, $p=0.003$, $\eta_p^2=0.488$). Post-hoc tests reveal that the value of pre-blank duration of 600ms and a post of 300ms was significantly different than all other data points ($p<0.001$ for all, except

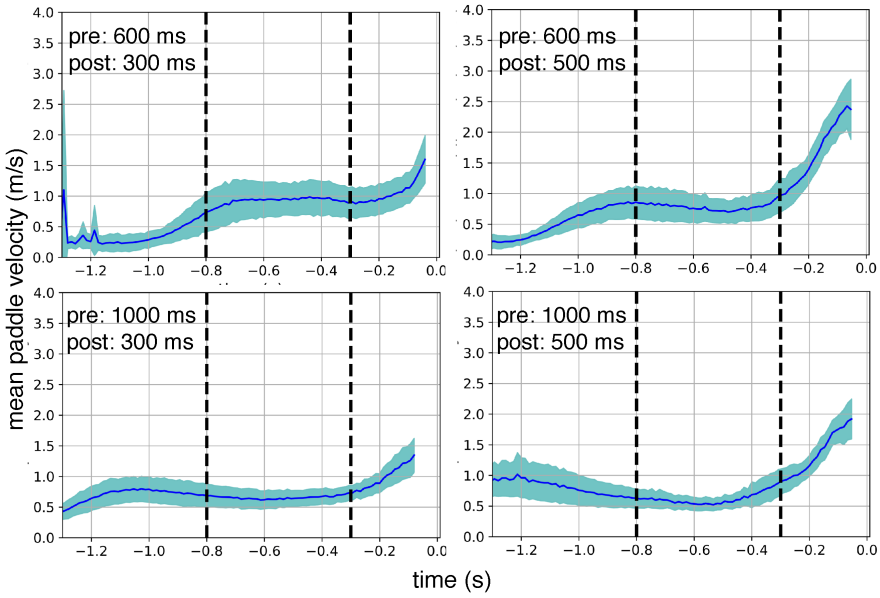


Figure 4.16: Paddle velocity relative to the ball's original time-to-contact. Vertical dashed lines indicate the onset/offset of the blank period. Shaded regions indicate 95% confidence intervals with between-subjects error removed.

($p<0.015$ pre/post of 800ms/300ms, assessed using Tukey's HSD). This effect suggests that, when subjects foresaw that a late blank period would leave little time for post-blank adjustments to paddle position, they initiated more dramatic movements of the paddle during the blank period.

Figure 4.17B presents the distance of the paddle, sampled at the time of the ball reappearance, from the ball's eventual passing location. Neither main effects nor their interaction were significant following Bonferroni correction(Pre: ($F(1.27,11.38)=5.329$, $p=0.034$, $\eta_p^2=0.372$, Post: $F(2,18)=4.41$, $p=0.028$, $\eta_p^2=0.329$, $F(4,36)=1.151$, $p=0.348$, $\eta_p^2=0.113$). This suggests that the parameters of the ball's trajectory did not affect the overall accuracy of the paddle's positioning at the time of ball reappearance.

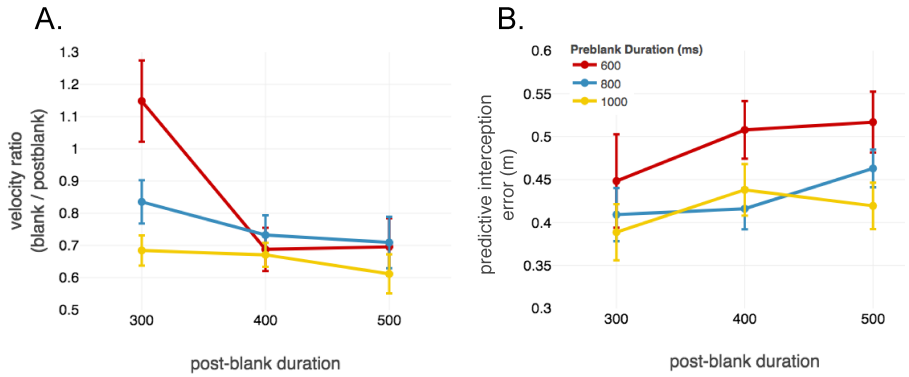


Figure 4.17: Measures of paddle positioning and movement timing indicate the relative role of predictive movements, and predictive placement of the paddle **A.** The trial-by-trial ratio of mean paddle velocity during the blank to the post-blank velocity. Values greater than one represent a higher mean paddle velocity during the blank period. **B.** Predictive interception error represents the distance from the paddle, sampled at the time of ball reappearance, to the ball's calculated passing point. Error bars represent 95% confidence intervals, with between-subject variability removed.

Summary of paddle placement and kinematics.

Analysis of paddle velocity during the blank and post-blank portions of the trial indicate that participants generally moved the paddle more quickly after ball reappearance than during the blank. The only time in which subjects moved the paddle more quickly during the blank was when the blank period occurred later in the trajectory (a pre-blank of 1000 ms, and a post-blank of 300 ms). Analysis of paddle positioning at the time of ball reappearance did not reveal an effect of post-blank duration on predictive paddle placement. Thus, analyses provide only weak support for the second hypothesis, that subjects will engage in more accurate prediction when the spatio-temporal demands of the task prevent online control.

	Model 1		Model 2	
	β	std. Error	β	std. Error
(Intercept)	0.46	(0.02)	0.62	(0.02)
visual prediction error (vpe)	0.03	(0.00)		
400 ms post	−0.04	(0.01)	−0.10	(0.02)
500 ms post	−0.07	(0.02)	−0.19	(0.03)
800 ms pre	−0.05	(0.01)	−0.04	(0.01)
1000 ms pre	−0.08	(0.01)	−0.05	(0.01)
vpe:400 ms post	−0.01	(0.00)		
vpe:500 ms post	−0.02	(0.00)		
pursuit gain (pg)			−0.13	(0.03)
pg: 400 ms post			0.07	(0.03)
pg: 500 ms post			0.11	(0.03)
Log Likelihood	696.11		636.74	
Num. obs.	1260		1260	
Num. groups: subject	10		10	
Var: subject (Intercept)	0.00		0.00	
Var: Residual	0.02		0.02	

Table 4.1: Effect of fixed and random effects for two models that included a random effect of subject, and fixed effects of pre-blank duration, post-blank duration, and the interaction of post-blank duration. In addition, Model 1 included the continuous predictor of visual prediction error (vpe), and model 2 pursuit gain (pg).

Predictive movements of gaze and the paddle covary.

A linear mixed model was used to investigate hypothesis 3: that when subjects are engaged in both visual and motor control, they will demonstrate correlated errors between the predictive placement of the eyes and paddle on a trial-by-trial basis. If so, this would be indicative of shared resources, such as a shared predictive representation. Model fits are overlaid upon the data in Figure 4.18, and model parameters are presented in Table 4.1.

Comparison of nested linear models fit to predictive interception error revealed a main effect of visual prediction error, (χ^2(1.00) = 570.69, p$<$ 0.001, dAIC = -58.49), post-blank duration (χ^2(2.00) = 650.07, p$<$ 0.001, dAIC = -154.75), pre-blank duration, (χ^2(2.00) = 680.92, p$<$ 0.001, dAIC = -57.71), and the interaction of post-blank duration with visual prediction error (χ^2(2.00) = 696.11, p$<$ 0.001, dAIC = -26.36). As such, all terms were included in the final model.

Although the quality of the model fit may vary by subject and condition, the trends indicated by model slopes are greatest at the post-blank level of 300 ms. In contrast, model slopes at 500 ms (right-most panels) indicate a weaker relationship between visual and motor prediction errors. These results support the hypothesis that visual-motor coordination is greatest when the ball reappears later in the ball's trajectory (at a post-blank duration of 300 ms).

A similar linear mixed model was used to investigate for a relationship between visual pursuit gain and predictive interception error, where both measures were sampled at the time of ball reappearance. Comparison of nested linear models revealed a main effect of pursuit gain (χ^2(1.00) = 570.69, p$<$ 0.001, dAIC = -58.49), post-blank duration (χ^2(2.00) = 615.72, p$<$ 0.001, dAIC = -86.05), pre-blank duration, (χ^2(2.00) = 629.34, p$<$ 0.001, dAIC = -23.24), and the interaction of post-blank duration with pursuit gain (χ^2(3.00) = 705.25, p$<$ 0.001, dAIC = -145.83). As such, all terms were included in the final model. Model fits are presented in Figure 4.19. Slopes were negative, indicating a drop in pursuit gain was accompanied by an increase in predictive interception error (Table 4.1). The higher gain values reflect lower ball speeds at ball reappearance and an

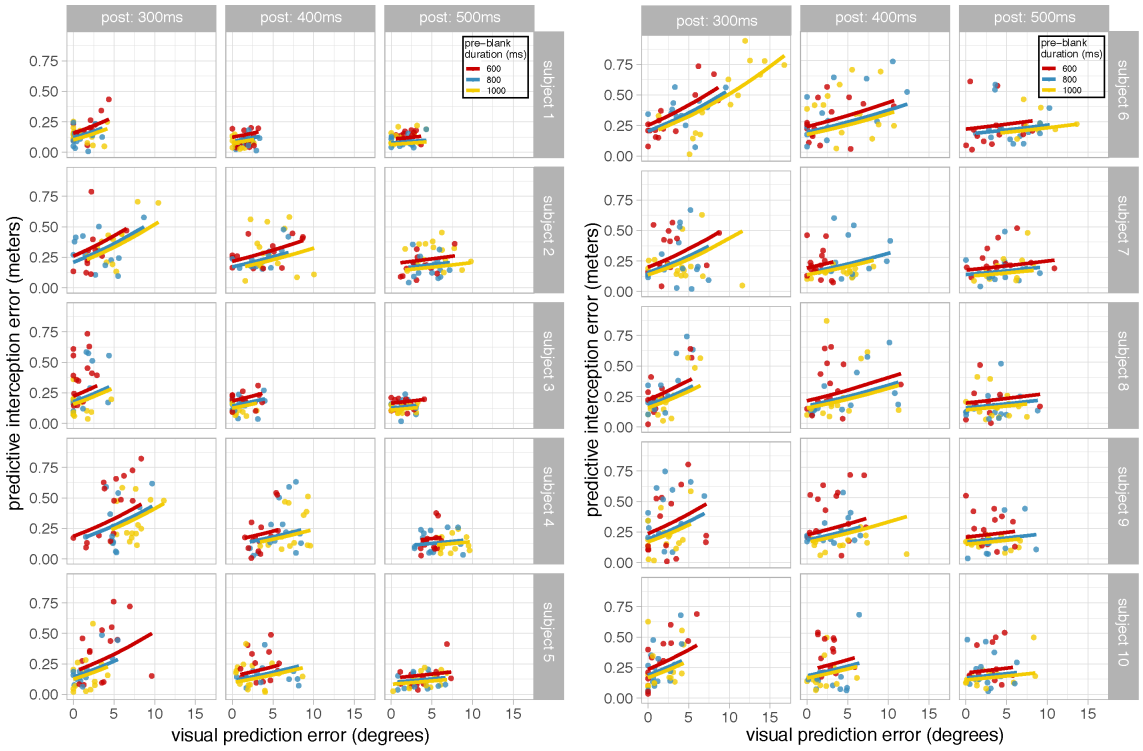


Figure 4.18: Each subplot presents the relationship between visual prediction error (abscissa), and predictive interception error (ordinate) by post-blank duration (column) for each subject (row). Each point represents a single trial. Colors represent pre-blank duration, and solid lines present the fit from the model presented in Equation4.1 for the range of actual values of visual prediction error.

increase in pursuit speed towards the end of the blank period.

4.2.4 Summary of hypotheses and results

Interception rates suggest that the task is most difficult at low values of post-blank duration, and that this effect cannot be attributed to the shape of the trajectory. An inspection of gaze behavior offers a possible explanation.

Consistent with hypothesis 1, subjects moved gaze in prediction of the ball's trajectory across

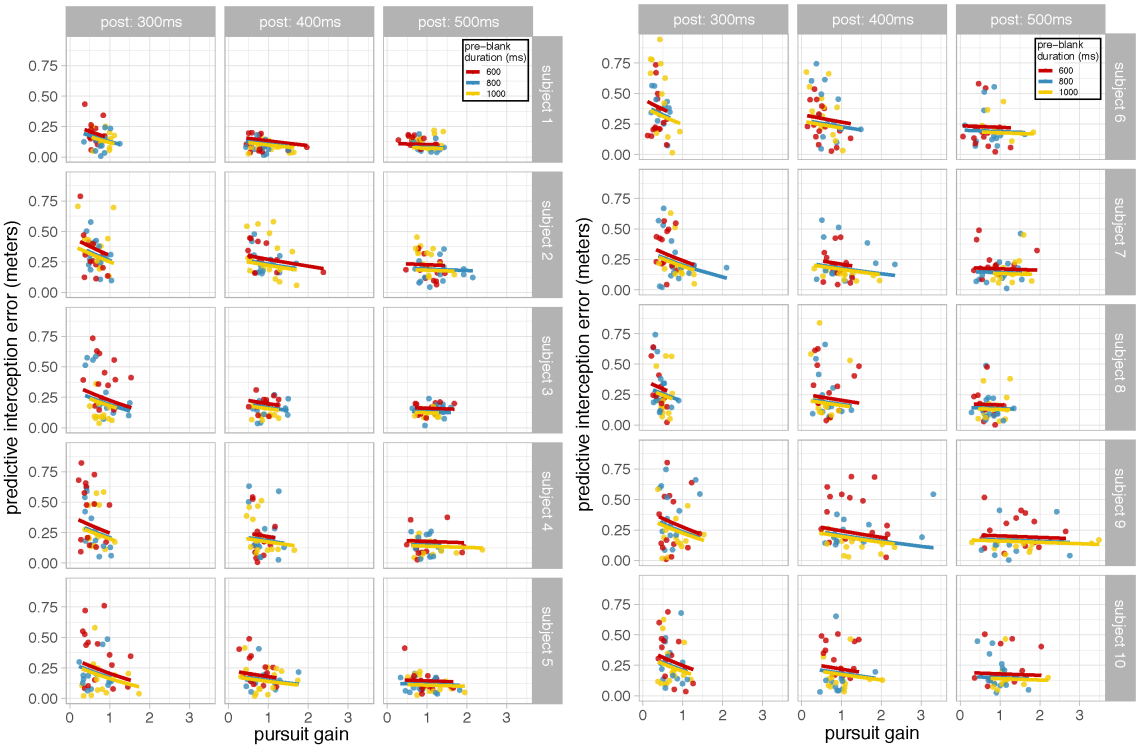


Figure 4.19: Each subplot presents the relationship between pursuit gain (abscissa), and interception error (ordinate) by post-blank duration (column) for each subject (row). Each point represents a single trial. Colors represent pre-blank duration, and solid lines present the fit from the model presented in Equation4.2 for the range of actual values of interception error.

the blank, in all conditions. The trajectory of gaze during the blank was curvilinear, appropriately scaled to the curvature of the ball as it moved around the head, and brought the gaze vector within 4 degrees of the ball upon reappearance, despite 10-13 degrees of ball movement. When the blank period occurred later in the trial, the ball would undergo greater displacement during the blank, and would be moving at a greater velocity upon reappearance. This resulted in a drop in pursuit gain that may explain the subject's difficulties in intercepting the ball.

Analysis of paddle kinematics revealed weak evidence for the second hypothesis, that subjects would switch to a more predictive mode of control at shorter values of post-blank duration, when

there was little time to resume online control after ball reappearance. This conclusion is based upon the findings that mid-blank adjustments to the paddle were greatest in magnitude when the blank occurred later in the trial.

The results supported hypothesis 3, that when participants were engaged in predictive control strategies for both gaze and the paddle (i.e. when post-blank duration was 300 ms), the two modalities would demonstrate correlated errors. Indeed, linear models reveal that visual prediction error and pursuit gain both covary with predictive interception error. This is indicative, but not conclusive evidence, for shared mechanisms behind the predictive control strategies guiding movements of gaze and the paddle.

4.2.5 General Conclusion & Discussion

This study provides valuable new insights into the role of prediction in naturalistic control of gaze and hand movements, and builds upon a growing body of literature suggesting that, although predictive control strategies may not be readily apparent in the analysis of navigation or effector placement [15], predictive mechanisms are fundamentally involved in the control of eye movements in visually guided action in natural or naturalistic contexts [59, 61, 62, 82, 83]. What is less clear, however, is how predictive gaze control is related to predictive control strategies for controlling the effector (e.g. the paddle). The present task was designed specifically to elicit prediction across both gaze and the effector, and as a result it provides new evidence concerning the relationship between the two modalities. Unsurprisingly, we found that gaze was consistently controlled in prediction of the ball's curvilinear trajectory through space while it was "blanked" for a brief duration leading up to an interception. Although predictive placement of the paddle was not mediated by the timing of the blank duration within the ball's trajectory, predictive control of paddle placement was apparent in the analysis of trial-by-trial covariations between the positioning of the gaze vector and the paddle: when the gaze vector accurately predicted the ball's

location at reappearance, the paddle was also held closer to the ball's eventual passing location. Because co-variation of gaze and paddle placement was present at the time of ball reappearance, it cannot be explained by the availability of post-blank visual information. Because this trial-by-trial co-variation was present within a single condition, its presence cannot be explained by changes in the timing of the blank. Together, these findings reveal that predictive strategies for paddle placement were more strongly mediated by the accuracy of gaze behavior than the range of variations in ball's trajectory adopted in this experiment.

What does the observed coupling between the predictive control of gaze and paddle position tell us about their underlying relationship? Because co-variation between gaze and paddle positioning was found when both modalities were sampled simultaneously, at the time of ball reappearance, it is not possible that the accurate visual prediction provided better (post-blank) visual information that causally led to better paddle placement. The alternative explanation is that the covariation is the result of a common cause, however the present data cannot distinguish between the possibility that the observed covariation reflects a shared upstream neural mechanism, or two largely independent mechanisms that rely upon common sources of visual information.

The data reveal that the strength of predictive covariation between the two modalities is mediated by the spatio-temporal demands of the task. This seems to suggest a transition from online to predictive modes of control that is mediated by post-blank duration, and thus presumably by the motor noise inherent in last-minute adjustments to paddle position. The reliability-mediated switching between online and predictive control strategies has also been demonstrated experimentally by coupling target disappearance to the initiation of movement in a 2D disc interception task [78] or in an interception task with real balls [77]. In both contexts, subjects had the choice to make a slow controlled movement to the predicted location of the invisible target, or to initiate movement late in the ball's trajectory, with increased motor noise. Subjects demonstrated a preference for a ballistic visually guided movement that maximized the duration of the visible portion of the ball's flight, and minimized the role of prediction. Thus, predictive strategies appear to be

adopted only when online control is prevented or degraded by experimental conditions. A similar proposal has been made by Belousov et al. [84], who modelled transitions between predictive and online control strategies when controlling a run to intercept a target, parameterized on the basis of observation noise, reaction time, and task duration.

The finding that interception rate is heavily influenced by the timing of the blank period contradicts a recent study by López-Moliner, which suggests no effect of the timing of an occlusion upon catching performance [85]. Gaze behavior was not measured. The difference can likely be attributed to differences in the types of trajectories used. In the study by López-Moliner, trajectories were thrown gently by the experimenter from a distance of about 75 cm to the approximately location of the catcher's hand, and the visual scene was entirely occluded for 250 ms of the ball's flight. In contrast, the trajectories experienced in our study spanned 20 m, balls were in the air for approximately 1.7 seconds, and passing location was randomized within a 1m x 1m plane. These differences suggest our task may have made it more difficult for subjects to predict the ball's final arrival location and, as a result, exaggerated the impact of the timing of the blank window upon interception performance. Because balls in our experiment were not thrown, subjects did not benefit from advanced visual information present in the movement of the thrower. Alternatively, differences in trajectory may have affected the visual information upon which subjects relied to predict the ball's future trajectory [86].

It is interesting to note that visual prediction may be more evident in data from subjects who had a more difficult time with the task. The role of individual differences is consistent in studies on visually guided behavior, and may provide a potential explanation for athletic expertise [87]. In fact, differences in eye movements have been found between individuals that vary in levels of expertise in multiple ball sports, including a simple interception task [63], cricket [59, 62], baseball [88], juggling [89, 90], table tennis [91], and tennis [92, 93]. Fooken et al. [73] found that the experience level of baseball players predicted skill level in a 2D target interception task in which subjects pointing at discs that disappeared while moving across the fronto-parallel plane.

Although the present study was not designed to address the factors contributing to the individual differences in gaze behavior and prediction, it is an area that may be addressed in the future.

4.3 Study3: An LSTM-RNN Model for Prediction

In the previous study, we showed subjects use prediction when the visual information of the ball is removed and the predictive strategies are apparent both in eye and hand movements. However, the mechanism of prediction and how predictive behavior is generated is a topic of debate. One solution for predictive control is to implement a full-blooded model of the kinematics and dynamics of a ball-in-flight and let the model use the last visual information to predict during the blank period. However, in many contexts, such as target interception and visual locomotion, emerging evidence disputes the necessity and also sufficiency of an internal model of the world [15, 61, 94]. In the following study, we present a parsimonious model of predictive ocular-motor behavior that uses the current and previous visual and proprioceptive information in order to guide the action in the future. This model allows us to investigate the effect of past information integration on the accuracy of prediction.

4.3.1 Statement of the Problem

In the 1960's, J.J. Gibson put forth his foundational set of theories concerning *ecological perception* and visually guided action [9]. Gibson theorized that the transformation from vision into action could be modeled as a closed-loop coupling between the parameters relevant to actions and the *optical variables* that forecast a task-relevant future world state. For example, when attempting to catch a ball in flight, one can couple the time of hand closure to the ball's time to contact, which is instantaneously specified throughout the ball's approach by optical variables such as optical τ [95–97], equal to the ratio of the ball's instantaneous optical angle over its rate of angular expansion. Although the existence of τ was originally thought to underly the perception of time-to-contact across a limited range of tasks and conditions, it has since been recognized that multiple, redundant optical variables are able to provide perceptual estimates of time-to-contact, but these sources vary in their reliability across task contexts [86, 98, 99]. The principles that determine the

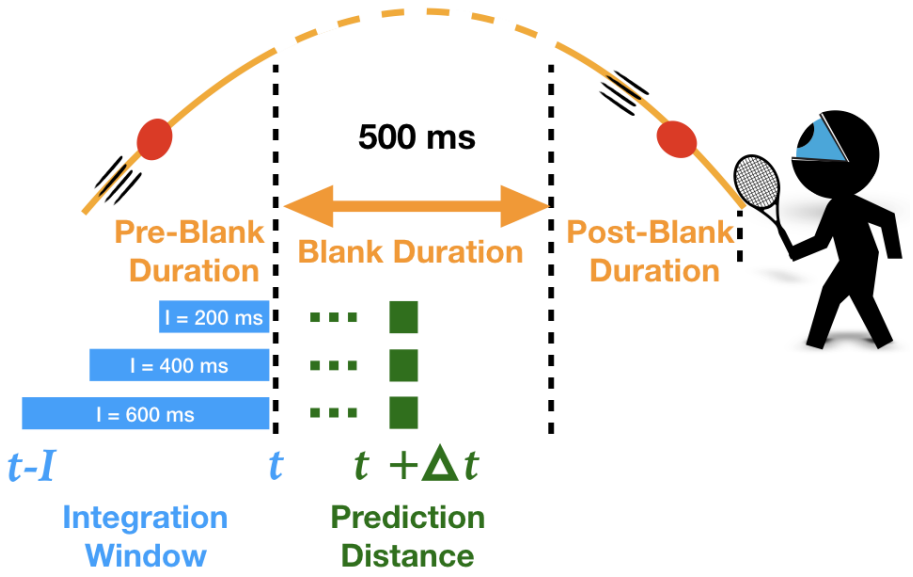


Figure 4.20: We developed a recurrent neural network model that reproduces human movements made in a virtual-reality ball catching task in which subjects must intercept a virtual ball that disappears for $500ms$ during flight. The model integrates visual and non-visual sources of information before a blank onset from time $t - I$ through time t, and uses this information to reproduce the subject behavior observed at time $t + \Delta t$. Multiple competing models were fit to the data for the purpose of exploring the minimum duration of pre-blank visual information (I) necessary to accurately reproduce behavior.

relative weightings placed by the perceptual system upon redundant optical variables that indicate a common task-related parameter remains a central question in the study of visually guided action. In this paper, we describe our modeling effort (as shown in Figure 4.20) to elucidate these principles.

Recently, it was demonstrated that the relative weightings placed upon these redundant optical variables by the perceptual system may be partially understood through the framework of maximum likelihood estimation (MLE), which is able to account for shifts in weighting upon variables within the course of a single action [100]. The authors found that reliable sources of optical information available early in the trial may influence behavior later in the trial if other reliable sources

do not present themselves. A notable advantage of the MLE framework is that perceptual estimates of the task-relevant parameter (e.g., time-to-contact) may be formed through the integration of information over extended periods of time, even if they are temporally distant from the time of motor output. Thus, the model is able to capture the well-known empirical observation that, in the presence of reliable sources of information, behavior is best characterized by an online coupling with negligible latencies [15], as well as the finding that humans are capable of accurate short-term predictions on the basis of previously observed visual information on the order of hundreds of milliseconds [61, 101]. Furthermore, there are evidence of predictive strategies aligned with ecological theory in other domains such as subjects tapping in synchrony with a metronome [102] or walking in groups [103] without the need to use and internal model. However, little is known about the parameters of temporal integration, whether there exist limits to the duration over which information may be integrated, or whether there are short-term limits to the temporal distance between the integration of information and its motor output.

This study aims to further characterize the temporal characteristics of information integration using a Virtual Reality (VR) system. A head mounted display equipped with a binocular eye tracker and motion-capture systems are used to record the gaze and movement behavior of participants placed in a virtual reality catching simulator in which visual information of the ball-in-flight is unavailable for 500 ms of its parabolic trajectory (the blank period, see Figure 4.20). To investigate the temporal limits over which visual and non-visual (e.g., kinesthetic) sources of information influence the motor output, we then train multiple models, each consisting of multiple recurrent neural networks (RNNs), to reproduce the gaze positions and hand movements observed during catching. Models vary in the duration of pre-blank visual information used to predict behavior during the blank (i.e. the integration duration), and subnetworks within a model vary in the temporal distance between the integration window and the motor output (i.e. prediction distance). The behavior of three representative models is shown in Figure 4.20. These models have integration durations of $I = \{200, 400, 600\}$ and they all predict the motor output at a particular time $t + \Delta t$.

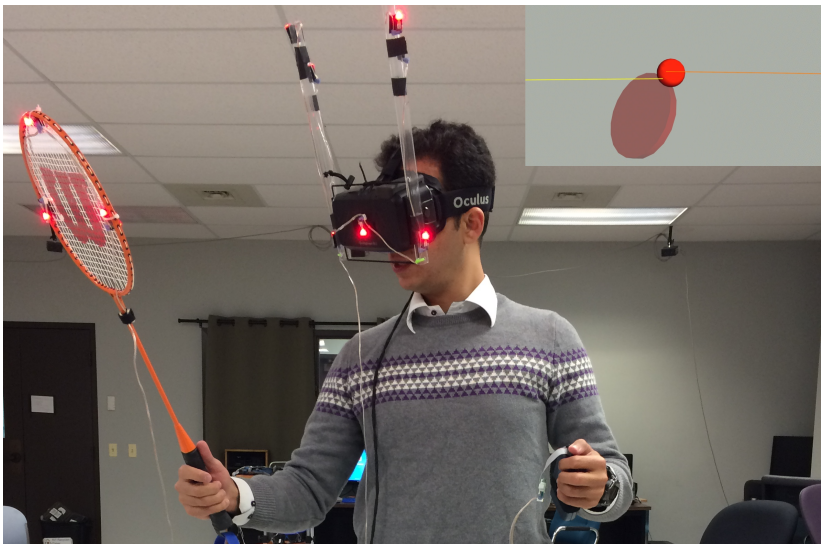


Figure 4.21: During data collection, subjects were immersed in a virtual ball catching task seen through an Oculus Rift DK2 with an integrated eye tracker. Movement was tracked using motion capture markers affixed to the head mounted display (HMD) and paddle. The scene inside the HMD is shown in the inset. The paddle was represented as a red disc, and gaze direction by yellow/orange vectors. These vectors were not visible to the observer at the time of data collection.

4.3.2 LSTM-RNN Model of Predictive Behavior

A single model of prediction across the blank duration consists of a group of long short-term memory (LSTM) subnetworks [104–108]. An LSTM-RNN is preferable to a simple RNN due to its robustness to exploding/vanishing gradient problems [105, 106]. A single model is presented in Figure 4.22 in which each row represents an individual subnetwork. The input to each subnetwork is a sequence of visual and non-visual sources of information observed within an integration window with an integration duration I. The right side of the integration window is always aligned with the last frame prior to the blanking of the ball at time t. This means that the integration window spans from time $t - I$ through time t. The integration duration I is constant across subnetworks belonging to a single model. The output of each subnetwork in the model is a discrete mapping from time t to time $t + \Delta t$, and the prediction distance Δt varies across subnetworks

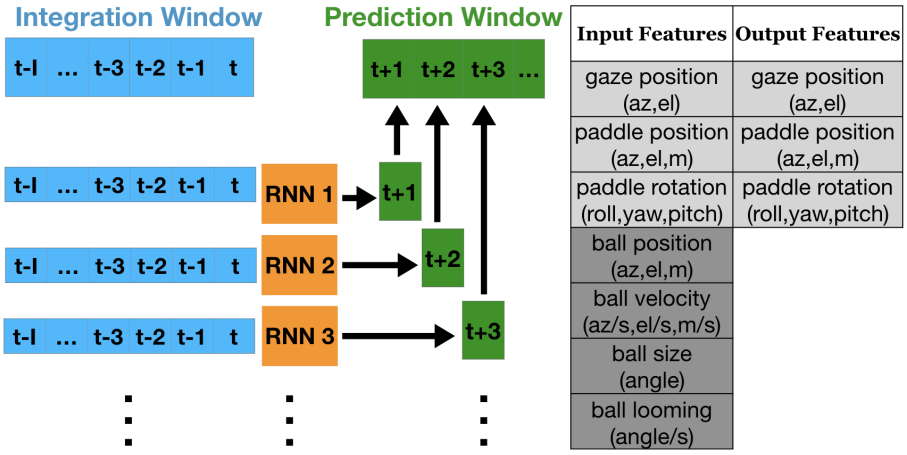


Figure 4.22: The left panel presents a single model consisting of collection of long short-term memory recurrent neural subnetworks, each of which is responsible for predicting the motor output at a single time step during the blank. The right panel presents the input & output feature vectors, along with their units.

in the model. Prediction across the blank period is facilitated by monotonically increasing Δt by frame-increments (13.3 ms) up to a duration equal to the blank period (500 ms).

4.3.3 Sub-network Inputs and Outputs

The input into each subnetwork consists of a 16-dimensional input feature vectors: the first 8 dimensions corresponding to the action/motor variables, and the remaining 8 dimensions corresponding to optical variables related to the ball movement (see Figure 4.22). Sources of information are directly calculated from the dataset geometry. The 16 element input feature vector corresponds to sources of visual information that are readily available from the stereoscopic imagery, and kinaesthetic information about the current state of the body (e.g. from proprioceptive systems, vestibular systems, and efference copy). Optical variables include ball angular position (degrees azimuth and elevation), velocity (degrees per second), ball depth from the head (meters), ball angular size (degrees), and the ball's angular rate of expansion (degrees/second). Information

from kinesthesis include paddle position (meters along X, Y, and Z), paddle rotation (Euler angles roll,pitch, and yaw) and angular gaze direction (degrees azimuth and elevation). All features were defined in the head-centered, egocentric coordinate system, with the up-vector aligned with gravity, and the horizontal vector parallel with the ground plane. To normalize the feature vectors, we subtracted the mean and divided each feature by its standard deviation, where the mean and standard deviation were computed using the entire training set. The model output is the predicted 8-dimensional action/motor state for each of the next Δt time steps, consisting of only position and orientation information.

4.3.4 Architecture, Training and Evaluation

Each of the LSTM subnetworks in the model has 1 hidden layer of 25 LSTM cells. In preliminary experiments, we did not observe improvements using additional cells. The hidden layer of each LSTM projects to a fully connected layer with 8 units that predict a future motor/action state. Because training was meant to account for predictive behavior, and not online control, we restricted training to periods in which the motor output occurred during the blank. Each model has 37 rows of subnetworks, hence each subnetwork is responsible for predicting the motor/action state at time Δt, where values range from 13.33 ms to 493.33 ms into the future, with a resolution of 13.33 ms.

We split the dataset into train (68%), validation (12%), and test (20%) partitions. The model was trained on all 135 trials of all 10 subjects. The models were optimized using the Adam optimizer with a learning rate of 0.0001 and the settings recommended in [109], e.g., batch size of 128 and 2000 epochs. Early stopping based on validation loss was also used with patience parameter set to 100. The dataset is formatted into Pandas data frame and is available as an online repository.

4.3.5 Training and Testing Results of the Models

Subjects on average caught the ball on 67% of trials (SD: 14%). During the blank period, the invisible ball moved between 10.3 degrees (pre-blank: 600, post-blank: 500 ms), and 12.6 degrees (pre-blank: 1000ms, post-blank: 300 ms) through the subject's visual field. During the blank, subjects tracked the ball through coordinated movements of the eyes and head. The ratio of angular displacement of gaze over that of the ball reveals that subjects accounted for 0.95 of the ball's displacement across all conditions (SD=0.11; t(9)=-1.43, p=0.187). Upon reappearance, the ball was moving approximately 34.1 degrees per second (SD: 4.3), and the gaze vector was well matched to the ball's angular velocity, as indicated by a pursuit gain (ratio of angular velocity of the ball over gaze) of 0.94 (SD: 0.11; t(9)=-1.14, p=0.28). There were also variations with the timing of the blank e.g., with variation of the pre-blank duration and post-blank duration.

4.3.6 Model performance

Figure 4.23 presents the mean-squared error (MSE) for four models ($I = \{27, 53, 200, \text{and } 600\}$ ms) when predicting gaze and motor behavior throughout the blank period. For reference, we also indicate the results of the linear regression between information prior to the blank period and the motor output. For all models, both MSE and variability increased with prediction distance. The four LSTM models outperformed the linear regression by a magnitude that grows with prediction distance. The observation that there is no added benefit to increasing the integration duration beyond 27 ms suggests that 27 ms of visual information prior the blank is sufficient to account for the predictive movements observed during the blank.

Figure 4.24 shows the root mean squared error for the model with I=27ms as a predictor of gaze position (Figure 4.24 top) and paddle position (Figure 4.24 bottom). In addition, dotted lines represent the the standard deviation of subject data around the grand mean, which is an estimate of the amount of unaccounted variance one would expect from a model that simply estimates the

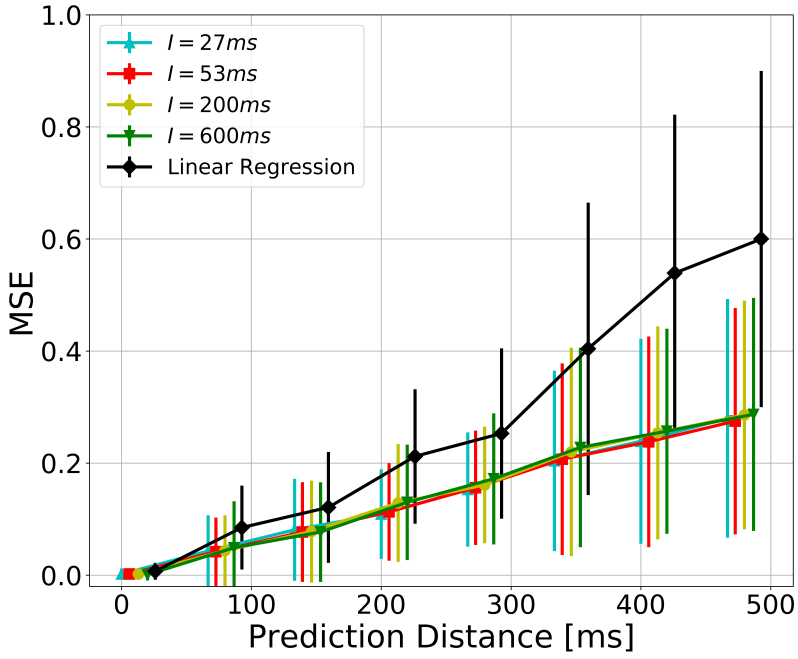


Figure 4.23: The performance of four models differing by integration duration ($I = \{27, 53, 200,$ and $600\}$ ms) as predictors of motor output throughout the $500ms$ blank period. For comparison, we also include a linear regression based upon the sensory evidence available before the blank.

mean value at each frame of the blank period. The observation that model RMSE is lower than this estimate is evidence that the model is able to account for trial-by-trial variations in ball trajectory on the basis of visual and kinesthetic input features.

4.3.7 Visual prediction, or a simple motor-to-motor mapping?

Although LSTM-based models of visual prediction outperformed linear regression as a predictor of gaze and motor behavior throughout the blank period, measurements of error alone cannot rule out the possibility that the model was performing a simple extrapolation of motor variables, while

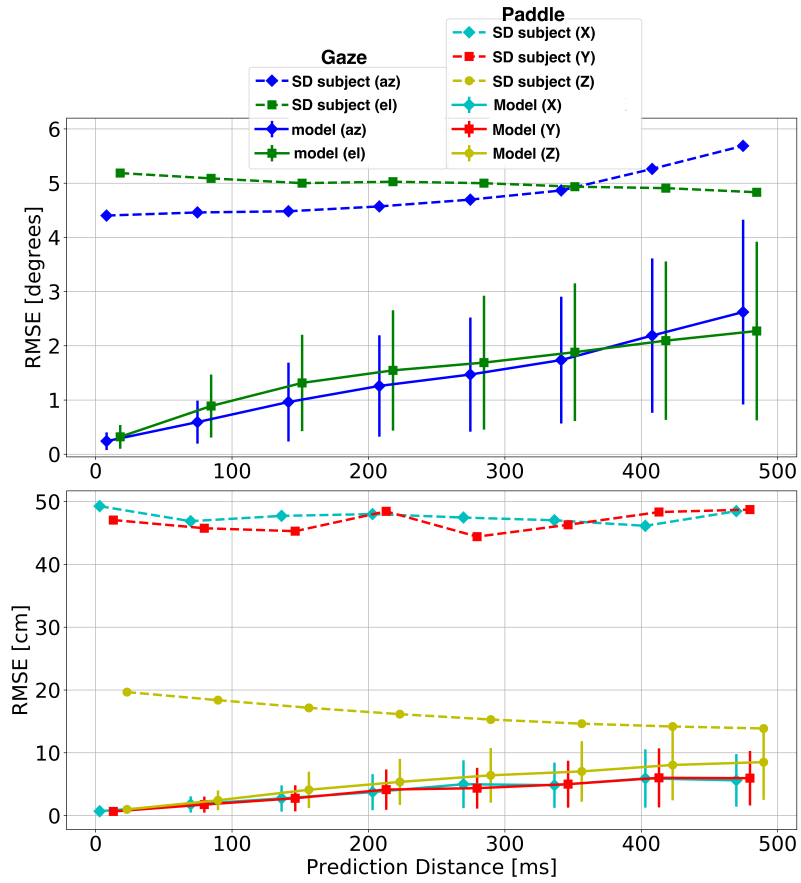


Figure 4.24: Root-mean squared error (RMSE) for the azimuth and elevation of the gaze vector in head-centered polar coordinates (top panel), and paddle position relative to the head in metric coordinates (bottom panel) for the model with $I = 27ms$. Dotted lines represent the standard deviation of the subject's gaze vector from the mean. These values provide an estimate of the RMSE expected for a model with an output equal to the per-frame mean gaze direction, and that does not account for trial-by-trial variations in the ball's trajectory.

disregarding the visual input state of the environment. To investigate, we ran a series of iterative tests in which individual input features were systematically removed from the subnetworks, and monitored the performance of subnetworks responsible for output at different stages of the blank period. The models/subnetwork training regime was not altered between tests, and included the full range of inputs. The assumption is that the ablation of an important input feature following training would result in an increase in mean reproduction error proportional to the feature's influence on the model's ability to reproduce the observed motor outputs. The results of these iterative ablation studies are presented in Figure 4.25 for two models $I = \{27, 600\}ms$, and for three prediction distances ($\Delta t = \{13, 267, 467\}ms$). To account for differences in units, the error values indicated by cell brightness were max-normalized across the output features represented by columns.

By comparing between rows within a single panel in Figure 4.25, one can visually compare the relative contribution of visual features (e.g. ball position, velocity, angular size, and looming) and features related to the subnetwork's motor output (e.g. gaze position, paddle position, paddle rotation) to movement reproduction. For example, in the bottom-left panel ($I = 600$ ms, $\Delta t =$ 13ms), it is clear that the subnetwork relied heavily upon visual sources of information concerning the state of the ball for the accurate reproduction of the observed motor outputs. Removing ball visual features caused on average 31% more error compared to gaze&paddle position/rotation. This suggests that, when integration time is long, visual information concerning the ball's trajectory is the best indicator of the motor behavior observed over short distances. There is a similar result when I=27, with the exception that the ablation of motor variables (the upper half of rows in Figure 4.25) degraded the reproduction of gaze elevation (column #1). The results of this ablation study suggest that, despite the lack of a benefit of increased integration time to overall RMSE, this may result in an increased robustness following the loss of an expected input feature.

Comparison across different prediction distances (between left, middle and right figures in Figure 4.23) suggests that, when predicting further in time, one must rely upon a combination of

input features related to the visual and motor state. This is true regardless of integration time, although values suggest a slight bias (less than 8%) towards motor variables when integration time is low (in the top middle and top right panels of Figure 4.25).

4.3.8 Discussion and Conclusions

In this study, we trained a series of competing models to reproduce the gaze and motor behavior made by subjects performing a catching task in which the virtual ball was transiently blanked for a portion of its flight. Under the constraints imposed by the task, only 27 ms of visual and kinesthetic information prior to the occlusion was necessary to accurately reproduce up to 500 ms of behavior following the removal of sensory feedback (Figure 4.23). Despite the low integration time of 27 ms, our models were able to predict gaze position within 3°of accuracy almost 500 ms after the ball's disappearance. This value is far below that expected by a model that simply estimates the time-varying mean, suggesting that the model was able to account for trial-by-trial behavior variations in response to changes in ball trajectory. Similarly, the model was able to reproduce hand position within 8.5 cm of error, 500 ms after the ball's disappearance (Figure 4.24). Although overall model performance did not vary with changes in integration duration, we found that the models ability to reproduce temporally distant behavior (i.e. at higher values of Δt) required input from both visual information and kinesthetic sources of information. The results of this task provide further insight into the temporal dynamics between sensory information and the motor output over the course of a single action.

The low-error observed for our model at prediction distances near to 500 ms provides evidence against the argument that accurate prediction requires internal models of physical dynamics (e.g. of Newton's law) for the continuous extrapolation of the ball's trajectory following occlusion [101]. Instead, our models learned temporally discrete mappings between evidence integrated from time $t-I$ through I, and a motor output at temporally discrete time in the future (time $t+\Delta t$).

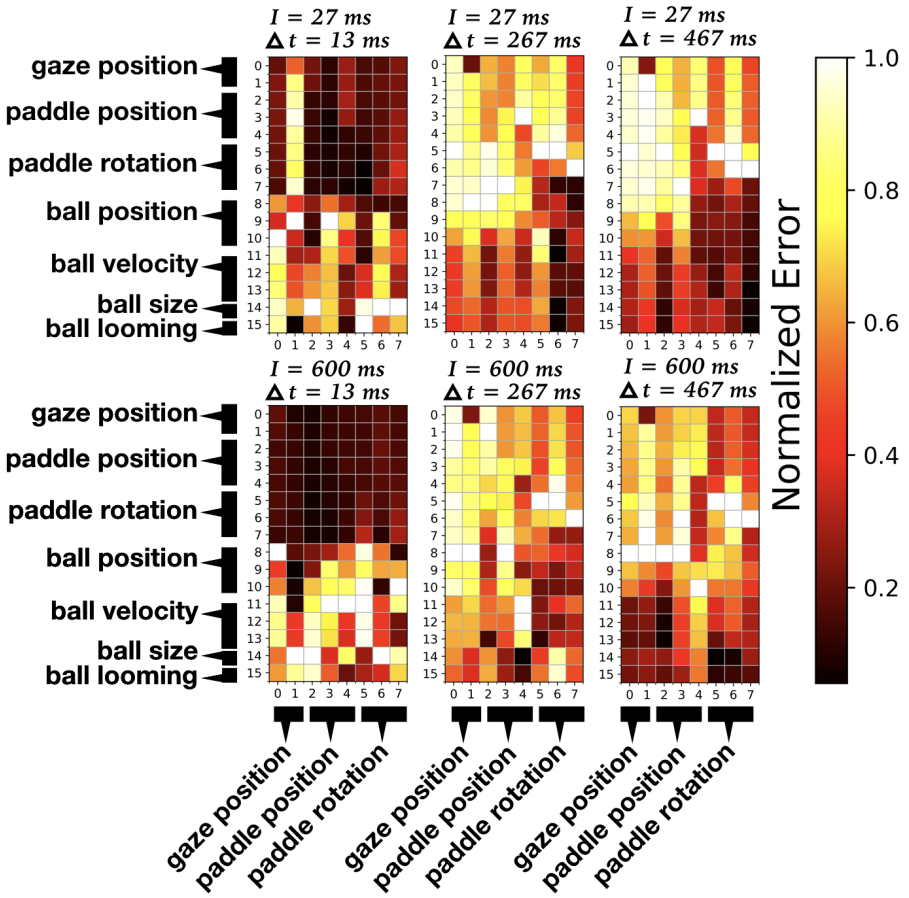


Figure 4.25: To test the relative contribution of input features to the accurate reproduction of the observed motor output, features were iteratively removed following training. Here, we present the resulting mean error in covariance following iterative input feature ablation for two values of $I = \{27, 600\}ms$ and three values of $\Delta t = \{13, 267, 467\}ms$. Rows indicate which feature was removed, columns correspond to the output feature, and brightness indicates the magnitude of error in covariance as a result of feature ablation.

The prediction distance at 500 ms in duration was roughly half what would be expected by a model that simply predicts the mean motor state (Figure 4.24), suggesting that such a simple mapping could be sufficient if one presupposes the availability of the information sources included in the model inputs. Moreover, it is notable that all variables were specified within a head-centered, egocentric reference frame, and did not presuppose reconstruction of the visual surroundings within a Euclidean frame of reference. Finally, by systematically exploring the error introduced by the ablation of input features in serial, we provided evidence against the possibility that the model was simply learning temporal correlations between subsequent motor states in the absence of visual input concerning ball position. Thus, the model serves as a proof-of-concept for the possibility that visual-motor prediction is a temporally discrete mapping between previously observed world states and a temporally distant motor output.

It is notable that the biological organism is subject to additional constraints not considered in the current architecture. Most notably, the model in no-way accounts for the influence of perceptual processing, or the perceptual sensitivities of the human visual system that further influence the reliability of information sources over time [110]. Similarly, the model does not account for short-term decay in visual working memory [111]. Finally, due to limitations in the LSTM-RNN framework preventing the use of dynamically sized integration durations within a single model, our model was unable to account for its own output motor states between the time of ball blanking (time t) and the current motor output (time $t + \Delta t$). Although our approach does demonstrate that 27 ms of sensory information is sufficient to explain predictive subject behavior, the human visual system must undoubtedly integrate across longer durations to overcome these biological constraints. The influence of these constraints might be explored in future work, for example, through systematic degradation of the visual input to reflect the constraints imposed by early visual processing.

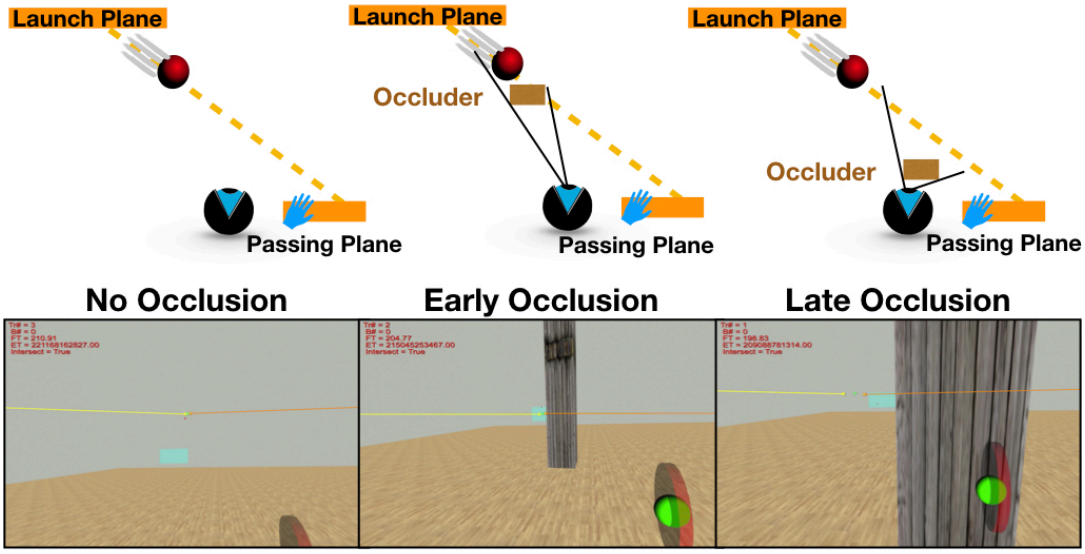
4.4 Study4: Prediction Explained by Inverse RL

In previous study we proposed a predictive model for ocular-motor performance during the time that the visual information is not available for our sensory system. Prediction is often treated as a separate mechanism that stands in for online control strategies only when necessary. In this study we investigate the transition from on-line to predictive strategies by manipulating time constraint and information reliability during a VR ball catching experiment. Here we propose characterization of human visual-motor strategies as a spectrum of behavior rather than two distinct notions. In this study we investigate the transition from on-line to predictive strategies in a VR ball catching experiment by manipulating ball speed, and the timing of a mid-flight occlusion of the ball in flight. Here we propose characterization of human visual-motor strategies as a spectrum of behavior rather than two distinct control strategies.

4.4.1 Statement of the Problem

Intercepting a ball in flight requires precise timing and accurate motor planning. Our sensory-processing system perceives relevant visual information and produces appropriate commands to drive our muscles for movement of the hand. Interception, and many other visually guided actions can be characterized as a closed loop coupling between a movement parameter and task-relevant sources of visual information referred to as *on-line* control [15, 24, 59, 62, 94]. More specifically, the feedback control loop receives visual information, produces the action, then uses the error signal to make corrections for the next time step [15, 21, 94]. As an example, one can visually track a moving target across the field of view by minimizing the angular distance of the position of the target upon the retina relative to the fovea [16].

However, there are several conditions in which visual-motor strategies based entirely on on-line control fail to explain the behavior. First, when there is a temporary loss of visual information, and second, when the delays in visual processing and transmitting the motor command prevent



	Early portion of trajectory	Late portion of trajectory
Estimate of POC	Less accurate	More accurate
Ocular Tracking	Easier (Lower angular velocity)	Harder (Higher angular velocity)
Optical Cue	Azimuthal & elevation angle	Target size & expansion rate
Hand Movement	Slower (more accurate)	Faster (less accurate)

Figure 4.26: Top-down view demonstration of different portions of a parabolic ball trajectory. During the first portion the estimates of TTC and POC relies mostly on azimuthal/elevation angle however during the late phase this estimate is mostly derived by target size and its expansion rate. Early occlusion forces the subjects to predict for information integration and late occlusion elicits predictive hand movement for interception. The yellow and orange horizontal lines are left and subject's right gaze vectors shown on experimenter monitor in order to check the eye tracker accuracy during calibration and throughout the experiment (not shown on HMD)

successful task execution [15, 94]. Therefore, humans switch to a predictive strategy when it's dictated by the quality of visual information or the temporal demands of the task [94]. Although the particular mechanism for prediction remains a topic of debate, it is clear that when visual information is removed, humans will orient behavior around a prediction of the world state a short time into the future. Zhao et al. [15] has categorized two classes of theory. *Model-based* predictive strategies rely upon an internal model of the movement dynamics, such as Newton's laws, to predict the future state, and this prediction is updated as the action unfolds. In contrast, *off-line* predictive strategies rely upon recent visual information about the ball trajectory to guide anticipatory action in the distant future. This predictive mechanism doesn't require an explicit representation of kinematics, rather it is based on a short-lived mapping between recent information and a future action.

Subjects may switch strategies even within a single task. For example, studies using eye tracking report that subjects commonly make predictive eye movements to the future position of a moving target [59, 61–63, 72], such as a ball right before it bounces [61], or when a moving target disappears mid-way through its parabolic trajectory [112][+cite our JOV]. This switch may be motivated by variation in the quality of visual information during task execution as differences in initial conditions affect the quality of visual information and temporal demands. For example, when estimating time-to-contact of a ball in flight, early estimates are less accurate than the later estimates that are based upon the final portion of the ball's trajectory [100, 113]. In contrast, when a ball is thrown from a shorter distance, the visual information about the last 200 ms of the ball trajectory has no effect on catching performance [85].

This study has been designed to investigate how the quality of information and the spatial and temporal demands of the task affect a transition from an on-line to predictive control strategy. Subjects were immersed in a virtual reality (VR) simulator and instructed to intercept a virtual ball using a paddle represented as a disc in the VR environment. To investigate the role of early vs late visual information on subjects' eye and hand movement strategies, an occluder was systematically

positioned to mask either the early or late portion of the 3D parabolic ball trajectory. In addition, to vary the temporal constraints of the task, balls approaches were consistent with either a shorter or longer time-of flight, yielding either a fast or slow velocity (see Figure 4.26).

We hypothesize that, in the late occlusion condition, subjects will adopt a predictive control strategy for placement of the paddle. This hypothesis is based upon the assumption that, because the late occlusion removes visual information from the final portion of the trajectory, subjects will not be incentivized to visually track the ball through the occlusion, for the reason that post-occlusion visual information is no longer used in guiding the final stages of the interception. As a result, participants will be forced to pre-program a paddle movement before the occlusion, on the basis of a predicted ball trajectory during the occlusion. In contrast, on the early occlusion trials, participants might adopt a strategy that prepares them for a return to on-line control after the occlusion duration. This hypothesis is motivated, in part, by studies showing predictive movement of the gaze to the future position of the ball after the bounce [61]. We predict that, in this task, a similar strategy would involve initiating predictive eye movements during the occlusion to the predicted location where the ball will reappear after occlusion. Finally, we hypothesize that this transition to a predictive control strategy should be most apparent in the fast ball condition, when there is less time for adjustment of gaze and hand positions during the trial.

4.4.2 Capturing Transitions Between Predictive and On-line Control Strategies

Although the on-line and predictive strategies are often characterized as two separate modes of control, it is more likely that control of action falls along a spectrum of behavior from fully on-line closed loop to off-line predictive strategies. To capture all possibilities along this spectrum, we use inverse reinforcement learning (IRL) framework to recover subjects' reward modules for different conditions. A reinforcement learner agent interacts with the environment by taking an action $a \in A$ i.e. moving the gaze or the hand. Based on this interaction it observes its own state

$s \in S$ and a reward (reinforcement) signal R provided by the environment [53, 54]. By repeating this so called exploration-exploitation loop, the agent gradually learns what is called an optimum policy $\pi^{\star}$. The learned optimum policy through experiencing a large combination of states and actions guarantees the maximum long-term reward for the agent. The quality of a state-action pair is often represented as a tuple $Q(s, a)$ to be maximized during training.

Inverse Reinforcement Learning Framework

The reward function of an RL agent provides a succinct definition of the agents strategy. However, in many practical circumstances, as when modeling human behavior, the rewarding mechanism is unknown thus creating a realistic model through forward RL framework is not feasible [76, 114–116]. Under certain conditions, assuming to have access to "expert" state-action pairs, inverse RL guarantees recovering the underlying reward function [55, 117, 118]. Here we use the linear programming implementation of IRL to recover the underlying reward values that best capture recorded gaze and hand movements demonstrated by subjects in the virtual reality interception task [118]. This algorithm is summarized into an optimization problem, where the heuristic in Eq.4.1 will pick the reward function from the solution set that maximizes the difference between the Qvalues of policy action and the second best action in all state space [55, 118]. We assume two reward modules for characterizing eye movements and a single reward module for driving hand movements. The actions associated with the gaze and hand movements are the same for each module [55, 76, 118].

$$\text{maximize} : \sum_{s \in \mathcal{S}} (Q^{\pi}(s, \pi(s)) - \max_{a \in \mathcal{A} \setminus \pi(s)} Q^{\pi}(s, a)) \qquad 4.1$$

State Space and Actions

Figure 4.27 shows the definition of actions and states for an agent responsible for movements of the gaze (e.g. the gaze agent), and an agent responsible for redirection of the paddle (the paddle

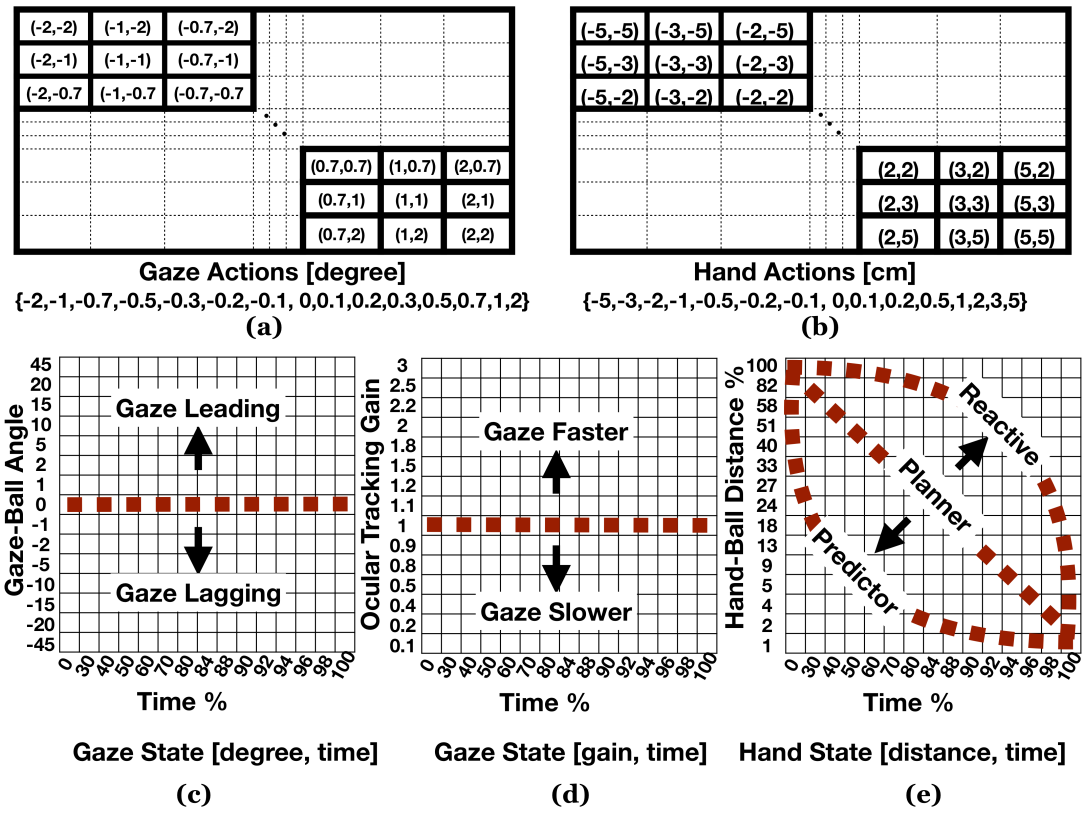
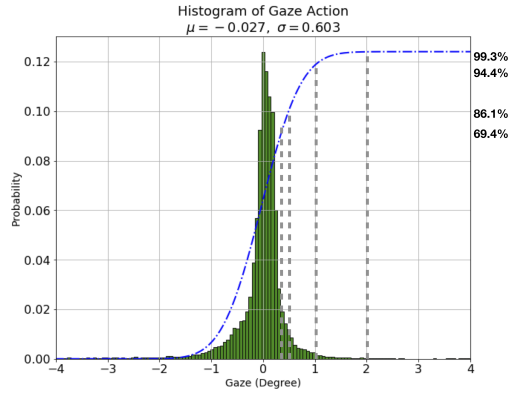
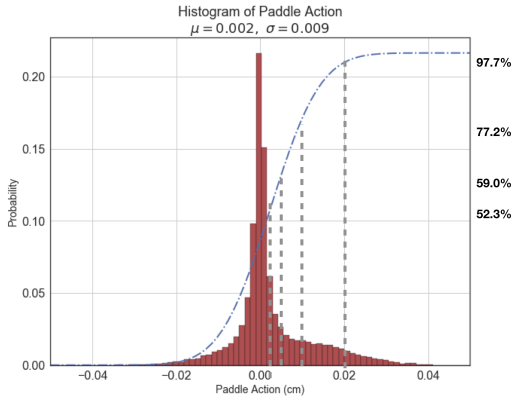


Figure 4.27: Top: The definition of (a) gaze actions and (b) hand actions. A grid of 15×15 actions available for gaze and hand agents selected independently that determines the magnitude of gaze/hand adjustment relative to previous time step. The grid sizes are selected based on distribution of subjects data. Bottom: Reward modules for (c) gaze-ball angle vs. time, (d) visual tracking gain which is the gaze-ball velocity ratio and (e) 2D hand distance to the position of contact where the ball hits the passing plane

agent). For gaze actions we assume a 15×15 grid that corresponds to gaze angular displacement from -2° to 2° in azimuth and elevation relative to the previous time step. The 15×15 state size was chosen because state spaces larger than 15x15 require more sophisticated techniques in order to guarantee convergence [55, 118]. Gaze and hand actions are shown in grid representation in Figure 4.27a and Figure 4.27b.



(a)



(b)

Figure 4.28: Distribution of (a)gaze and (b) paddle movements between two consecutive frames. The cumulative distribution of the data (blue dashed lines) show what percentage of the distribution is explained

In order to efficiently sample from the subjects' actions we incrementally increase the grid (bin) size according to the statistics of the movement. The distributions of gaze and paddle movements between two consecutive frames are shown in 4.28. The cumulative distribution function (cdf) is plotted using a blue dashed line. The grid size for each bin is chosen so that they contain approximately similar percentages of the data. The goal here is to assign smaller bins to the portions of the distribution that covers most of the data. For example, in Figure 4.28 (a) 69.4% of

the data lies between [-0.5°, 0.5°], where we assign 7 bins each of them containing 6.9% of the distribution. Similarly for the range of [-1°,1°] the added 16% falls into 4 bins each containing 4% of the data and finally the remaining 8% of the distribution falls into two bins each capturing 4% of the total data. Two bins covered the tails of the distribution ($\Delta_{gaze} < -2°$, and $\Delta_{gaze} > 2°$). We used the same technique in order to define the gride sizes for hand actions. The values reported in Table 4.2 & Table 4.3 report the range, total percentage, bin number and the percentage of the data that falls into each bin for the gaze and hand agent.

Table 4.2: Distribution of Data for Gaze Action Grids

Range (°)	Total%	Number of Bins	%Each Bin
[-0.3, 0.3]	69.4%	7	9.9%
[-0.5, 0.5]	16.7%	4	4.1%
[-1, 1]	8.3%	2	4.1%
[-2, 2]	5.6%	2	2.3%

Table 4.3: Distribution of Data for Hand Action Grids

Range (cm)	Total%	Number of Bins	%Each Bin
[-0.2, 0.2]	59.0%	7	8.4%
[-0.5, 0.5]	18.2%	2	9.1%
[-3, 3]	20.5%	4	5.1%
[-5, 5]	2.3%	2	1.1%

The gaze agent will update its state by taking actions that maximizes the two reward modules shown in Figure 4.27c and Figure 4.27d. In order to capture the agent states, a similar methodology was used to determine bin size when discretizing the agent states of gaze-ball angle and gaze-ball tracking gain vs normalized flight time. table 4.2 & 4.3 for gaze-ball angle and gaze-ball tracking gain vs normalized flight time (in percentage). At every step the normalized flight time along with angular distance between gaze and ball determines gaze state as shown in Figure 4.27c. Also the ratio of gaze over ball velocity determines the gain module state as shown in Figure 4.27(d). If the

subjects were perfectly tracking the ball with their gaze, their recovered reward values explaining their behavior would look like the dashed red lines in Figure 4.27c and Figure 4.27d. However, if their gaze position was leading/lagging or moving slower/faster than the ball, their reward values would have been above/below the dashed red line. Therefore the definition of reward modules for visual tracking the target allow us to characterize a range of strategies that the subjects could have adopted. Note that, using the same method discussed above, the temporal axis of the state representation is dynamically sampled according to the timing parameters of the task (occlusion on/off set) that will be discussed further in the results section. For example, the first 30% of the flight time does not include any significant event other than subjects engaging in visually tracking the ball (the first ≈300 ms for the fast and ≈400 ms for the slow ball trajectory). Therefore this range is assigned to one bin in Figure 4.27c,d & e. On the other hand the last portion of the flight time (≈150-200 ms) is where we'd expect to see larger differences in subject strategies for hand/gaze movement. Therefore the last portion of the flight time is represented using smaller grid size. These timings are explained in more detail in Figure 4.30 using the vertical dashed lines and the arrows signifying the low and high angular velocity of the ball.

For movement of the paddle, the state representation is the 2D distance between the paddle and ideal position of contact (where the ball hits the passing plane) vs normalized flight time. For simplicity, here we assume that there is a transformation from the exocentric to egocentric frame of reference that allows the paddle agent to observe its state as a measure of instantaneous 2D distance between the paddle and the ball final position.

The red dashed lines for the paddle state show three possible paddle movement strategies. The predictive strategy reduces the distance earlier in the trial when compared to the strategy of a reactive subject that waits until last minute to move the paddle, as shown by the dashed lines in Figure 4.27e. Since the normalized time is represented on the horizontal axis of the state modules, offline strategies when the ball was occluded will be compared to the strategies when the ball was visible [61, 112].

4.4.3 Experiment Design & Data Collection

The twenty three right handed participants (17 male, 6 female) were between 19-33 years of age and had normal vision in the absence of visual correction. The subjects participated in the experiment were equally credited through course management system aligned with codes of ethics. This study was approved by the institutional review board at the Rochester Institute of Technology.

Apparatus

Stimuli were delivered by an Intel i7-based PC with an NVIDIA GTX 690 connected to the Oculus Rift DK2 head mounted display (HMD), and an NVIDIA GTX 760 connected to the experimenter's desktop display. The computer ran Windows 7, and the virtual environment was rendered using the Vizard Virtual Reality toolkit by *Worldviz*. Physics were simulated using the OpenODE physics engine so that ball trajectories matched those expected within a real-world environment in the absence of wind resistance (Figure 4.26). The Oculus Rift DK2 HMD has an approximate field of view of 100°, and an approximate angular resolution of 9-12 pixels per degree, depending on the position of the eye inside the HMD [119]. Head and paddle position/orientation were recorded at 75 Hz using a 14 camera *Phasespace* X2 motion capture system, with a measured latency of between the time a sensed movement would be reflected on the display of less than 30 ms. Eye movements were recorded with an *SMI* binocular eye tracker running at 75 Hz. A post-hoc correction was applied, as described in [48], to correct for helmet slippage and other sources of spatial error in the eye tracking data. The average eye tracking accuracy after calibration and correction was 0.53° for the central visual field ($FOV < 10^{\circ}$) and 2.51° in the periphery ($10^{\circ} < FOV < 30^{\circ}$).

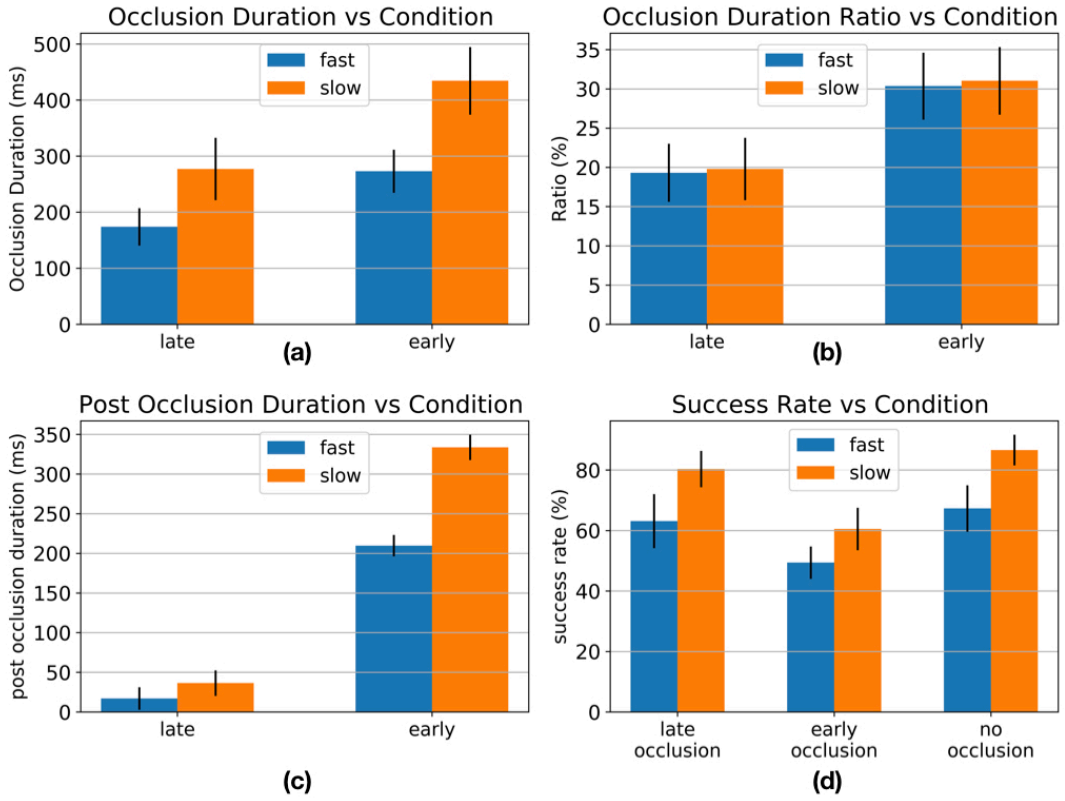


Figure 4.29: Timing parameters of the experiment design as the result of physical confounds of 3D ball trajectory in head centered angular coordinate system (a) Occlusion duration for the fast vs slow and the late vs early conditions (b) ratio of occlusion duration over flight time determining what percentage of the flight duration the ball was occluded (c) post occlusion duration for fast vs slow and the late vs early conditions. The error bars represent the between subject deviation of occlusion times due to lateral movements of the subjects head during the experiment. (d) success rate for all six conditions. The error bars represent between subject deviation from the mean

Procedure, Conditions and Confounds

The red virtual ball was launched from plane shown in cyan in Figure 4.26 that was 4 m wide $\times$ 1.5 m high and parallel to the virtual room's X-axis to a passing location randomly selected from a 1.5 m $\times$ 1.5 m plane near the subject, also parallel to the room's X-axis. During data collection subjects were standing in a green transparent standing box (not shown in the figure).

To launch the virtual ball subjects were required to hold the paddle (red disc shown in Figure 4.26) such that its center intersected with a green sphere that designated the required initial paddle position as shown in Figure 4.26. Once the paddle was appropriately positioned, the experimenter would initiate the launch of the ball. Then both the standing box and the green sphere would disappear, signaling the start of a ball catching trial. If subjects moved the paddle before the launch, the trial was terminated. After the ball was launched the subject was free to move his head and hand to intercept the virtual ball. If the ball was successfully intercepted, it would stick to the red virtual paddle and a collision sound was generated. If the subjects missed the ball it would hit the passing plane on their right side and a beep sound was generated to provide them with visual and auditory feedback.

Two ball flight times of 900 and 1400 ms were selected for the fast and slow conditions, respectively. For each flight time, before the ball was launched initial and final positions were randomly selected from the launch and passing planes. On occlusion trials, a vertical wooden box was placed either 2.5 meters (early occlusion) or 8 meters (late occlusion) down the axis spanning the room's depth relative the participant's standing location. No occluder was rendered on trials belonging to the no-occlusion condition. Once the occluder distance was selected, the lateral position of the occluder was adjusted to generate a preselected occlusion duration. This adjustment was based on the position of the standing box and the time of flight (see the top row of Figure 4.26). Each subject performed 150 ball catching trials consisted of 25 repetitions of each six conditions $2(ball \quad velocity) \times 3(occlusion \quad types) \times 25(repititions) = 150$ trials.

Since the time and duration of the ball's passage behind the occluder is influenced by small lateral movements of the subject's head, the mean and standard deviation in measured occlusion and post occlusion duration values are shown in Figure 4.29a-c. The time of occlusion was measured by casting a ray from the mid-point between the left and right eye (the cyclopean eye node) to the position of the virtual ball. From cyclopean eye node (the point between right and left eye) a 3D vector was generated to the position of the virtual ball. To measure the precise timing of when the ball goes behind the occluder and when it reappears, the collision of this vector with the occluder object was detected and recorded in the data collection pipeline. The results of this analysis are presented in Figure 4.29, which reveals that the occlusion duration was shortest for the late occlusion condition, when the occlusion occurs closer to the subject's head, and thus the ball's angular velocity is higher.

Data Preprocessing & Eye-Hand Movement Analysis

The subjects 3D gaze vector is calculated by averaging the left/right eye-in-head vectors provided by the SMI eye tracker to produce a single unit vector that represents the cyclopean gaze direction. The transformation matrix of the motion-tracked head was applied to the vector to cast its appropriate location in front of the head in the virtual world. Data was passed through a median filter (length 3). Here we assess the quality of visually tracking the ball which is a combination of smooth pursuit and intermittent catch-up saccades.

Three metrics are calculated in angular space or spherical coordinate system. First, the angle between 3D gaze vector and head-to-ball vector that determines the instantaneous visual tracking error in degrees. This metric is referred to as the gaze-to-ball angle. This angle is composed of an azimuthal and elevation angle. Second, the direction of the gaze vector movement is compared to the direction of the ball movement using the projection of the two 3D vectors on to the XY plane. For example if the ball or gaze was moving from left to right on a straight line the direction angle would be zero degrees and if they were moving from right to left it would be 180 degrees. The

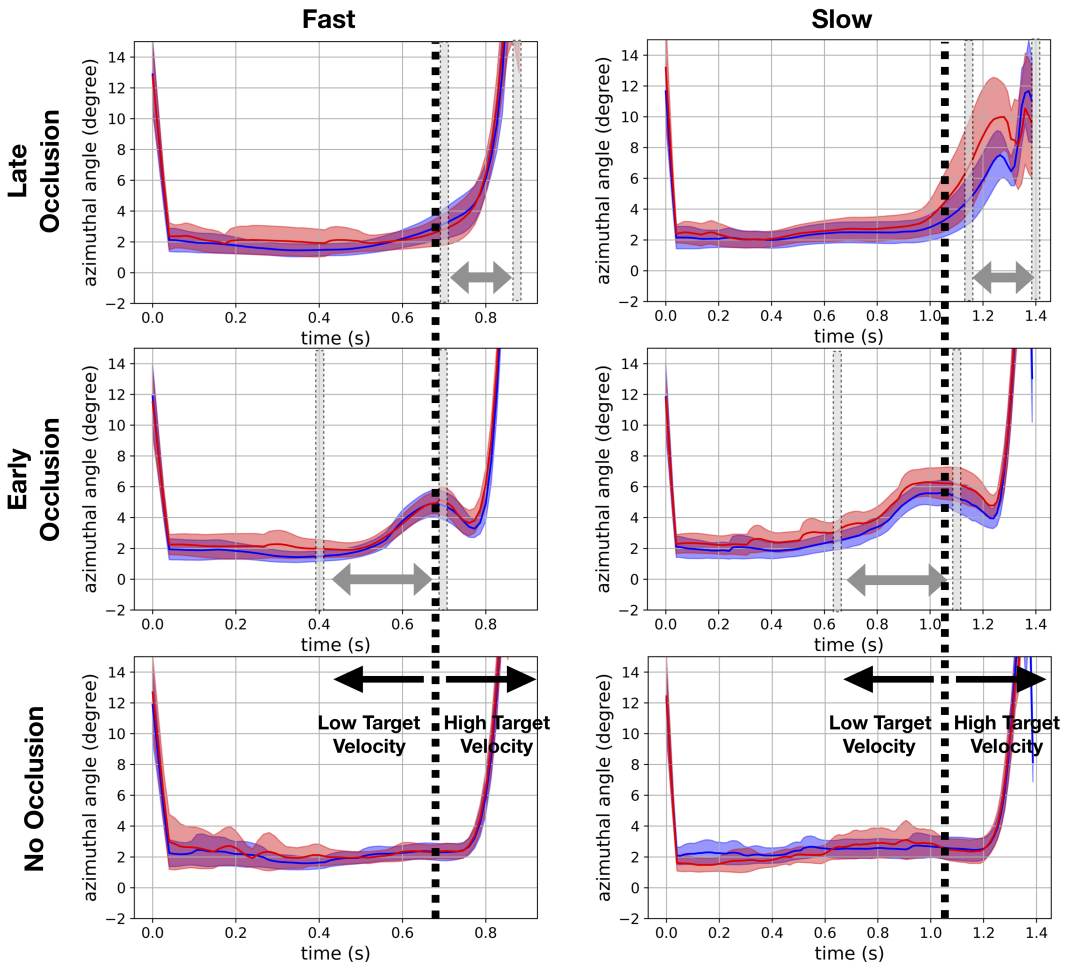


Figure 4.30: Mean angular distance between the gaze vector and the ball over the course of each trial type. Colored areas indicate 95% confidence intervals. Successful and failed trials are shown in blue and red respectively. Vertical gray lines indicate the start and end of occlusion duration. Black dashed lines along side the vectors show the window of slow vs high ball velocity for analysis

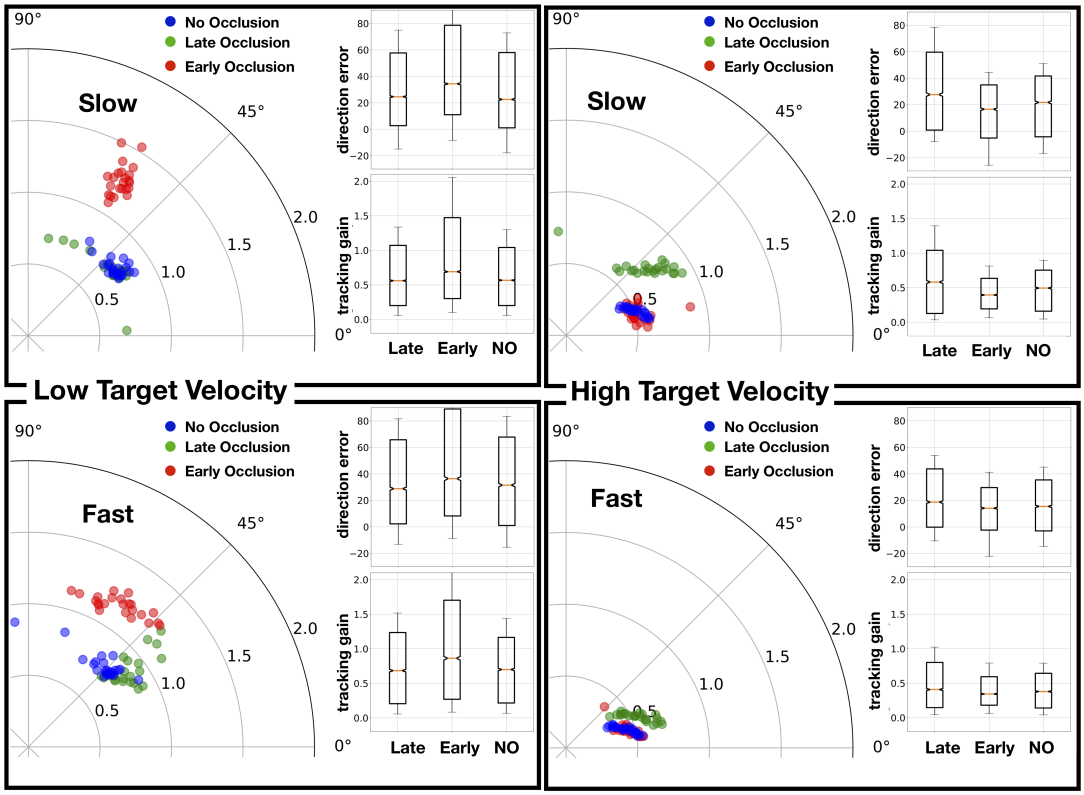


Figure 4.31: Distribution of mean gaze-ball direction error vs visual tracking gain for two occlusion conditions measured during planning and action phase. Each point represents one subjects mean value and the occlusion conditions are presented in color. Blue shows baseline performance (no occlusion), green shows late and red shows early occlusion conditions

difference between these two angles determines the accuracy of visually tracking the ball in terms of direction. Finally, we calculate the ratio of instantaneous velocity of the gaze or head-to-ball vector as another measure of successful visual tracking. This ratio is referred to as visual tracking gain and it would be close to one if the gaze angular velocity matches the ball velocity.

4.4.4 Results

Success Rate

To assess the degree with which early vs. late occlusions affected the subject's ability to catch the ball, we measured their overall catching rate. Figure 4.29d shows the success rate for all subjects for different conditions. The first observation is that when the ball was occluded early in the trial, although subjects were less successful than the late or no-occlusion conditions. In contrast, there was no effect of a late occlusion upon catching rate . The reason for this could be that the information available during the early portion of the trial is more important than the late information. In contrast, the late occlusion condition did not differ from the no-occlusion condition. The effect of the early, but not late occlusion, suggests that the information available during the early portion of the trial is more important than the late information.

This suggests that removing the late visual information about the ball trajectory does not significantly affect motor planning, and that the hand movement was planned before the time that the ball passed behind the occluder, on the basis of a predicted ball trajectory. This finding is inconsistent with the findings reported by [85], which found that subjects had difficulty catching when vision was occluded for the final 200 ms of ball flight. One possibility is that this is due to the nature of the task. Whereas their study required a precise grasp of the ball-in-flight, the present task required only collision between the ball and virtual paddle for a successful interception.

In all three occlusion conditions, subjects were more successful catching the slower flying balls. This could be due to the fact that, when tracking a fast moving ball, planning and executing the right paddle movement is harder for faster moving balls, and possibly involves a switch from a more reactive to a more predictive mode of control.

In summary, subjects had the most difficulty catching the ball when they could not integrate information about the early portion of the ball trajectory, and when the temporal demands of the task were highest. In the following sections, we will investigate gaze data and paddle position-

ing for evidence that these are also the conditions under which participants switched to a more predictive mode of control.

Gaze-to-ball Angle

Figure 4.30 shows the azimuthal gaze-to-ball angle vs time for all 6 conditions and all subjects. As it is shown, when there was no occlusion (the baseline performance), subjects engaged with tracking the ball approximately 100 ms after the launch. They maintained a low gaze-to-ball angle (below 4 degrees) until the last 200 ms, at which point the angle increased to values above 15 degrees prior to the the ball's collision with the paddle, or the passing plane. This rapid increase in the gaze-to-ball angle in the final stages of the ball's approach is likely due to angular acceleration of the ball around the head, and the tight relationship between the ball's distance from the head, and the speed of its movement through the visual field. The general trend of this sudden increase is observed both for slow and fast ball trajectories. During an early occlusion, subjects consistently moved their gaze to the predicted location of the ball's reappearance from behind the occluder, resulting in a temporary increase in the instantaneous gaze-to-ball angle. This likely facilitated a rapid return to visual tracking of the ball after occlusion (see Figure 4.30 middle row).

The gaze-to-ball angle on late occlusion trials follows the same general pattern as in the no-occlusion condition up until the ball passes behind the occluder (see Figure 4.30 top row). However, just as in the early occlusion conditions, subjects in the slow, late occlusion condition made mid-occlusion eye movements to the edge of the occluder in prediction of the ball's location of reappearance. One possible explanation is, since the ball is moving more slowly through the subject's visual field, subjects attempt to re-engage visual tracking of the ball after it reappears.

In summary, visual inspection of the average diagram of gaze-to-ball angles over time for each condition reveals that, in the early occlusion and the slow-late occlusion conditions, visual occlusion of the ball elicits a predictive eye movement to the future position of the ball after reappearance. This behavior is aligned with the predictive eye movement observed right before

the bounce in a similar interception task [61].

Visual Tracking Quality

In order to track the ball or to predict its trajectory during occlusion, beside angular error, subjects also need to take into account the angular velocity of the ball and its direction of movement. As shown in Figure 4.30 subjects tend to have 2-3 degrees of visual tracking error(gaze-to-ball angle) and this is larger than the size of the fovea (about 2 degrees) [120]. Furthermore toward the end of ball trajectory the angular velocity of the ball increases exponentially that makes it much harder to be tracked due to inherent motor delay in our head-eye movements. Therefore, instantaneous gaze-to-ball angle alone is not a sufficient representation of subject's visual tracking quality.

Figure 4.31 presents subjects eye movement statistics using two metrics based on direction of movement and visual tracking gain. The angles in the polar plot show the difference between gaze and ball movement direction. The radial distance from the center of the polar plot presents the visual tracking gain. Therefore, a perfect visual tracking strategy would have values close to $(r, \theta) = (1, 0^\circ)$.

The left column of Figure 4.31 represents each subject's average performance during the early phase which is the average timing of occlusion in the early condition (460-660 ms for the fast condition and 660-1100 ms for the slow condition). Similarly the right column of Figure 4.31 shows the result for the late phase (660-900 ms for the fast condition and 1100-1400 ms for the slow condition). Each point represents the mean value for each subject and color coded as blue:no occlusion, green:late occlusion and red:early occlusion. The left column of Figure 4.31 shows that subjects increase the visual tracking gain during early occlusion while in the other two conditions (late and no occlusion shown in green and blue) their visual tracking performance remains the same . The clustered blue and green points compared to the distant red cluster reveals the difference in subjects visual tracking strategy during early occlusion compared to the baseline condition. This period is aligned with the bump in Figure 4.30 middle row as described earlier.s

In summary, for both slow and fast ball trajectories, subjects increase their visual tracking gain during occlusion, as they redirect gaze to the predicted location of ball reappearance.

During the late phase, red and blue points are clustered around each other, however, the green points representing late occlusion are not. It is important to note that during the late phase, the visual tracking gain is below one for all subjects, mainly due to the high angular velocity/acceleration of the ball during the late portion of the trial. Using a similar comparison, even though it is much harder to keep up with the target velocity, subjects use higher visual tracking gain during late occlusion compared to baseline performance.

In summary, the analysis of tracking direction error and tracking gain reveals that, regardless of condition, subjects change their visual tracking strategy when the ball is occluded. One explanation for this predictive behavior could be that, leading up to a foreseeable occlusion of the ball, subjects optimize behavior for a quick return to on-line tracking of the ball after its reappearance. It is surprising, however, that subjects adopted this strategy in the late occlusion condition, when the short post-occlusion duration prohibited a return to on-line control. One possibility is that maintaining accurate visual tracking is somehow beneficial to guiding the manual interception. Alternatively, it may be that these subjects are looking ahead, and towards the location of a potential collision of the ball with the paddle.

Absolute Paddle Velocity

To investigate for an influence of occlusion timing upon movement of the paddle, movement strategies on the early/late occlusion trials can be compared to the no-occlusion condition. Here, we analyze absolute paddle velocity through the 3D environment to determine whether subjects' motor planning and execution indicative of a predictive or reactive control strategy. Paddle movement initiation started at approximately 400 ms after the launch, for all conditions. Since the subjects were instructed to hold the paddle at the initial position, and ball final position was randomly selected from a 2D Gaussian distribution, the direction and magnitude of paddle movement required

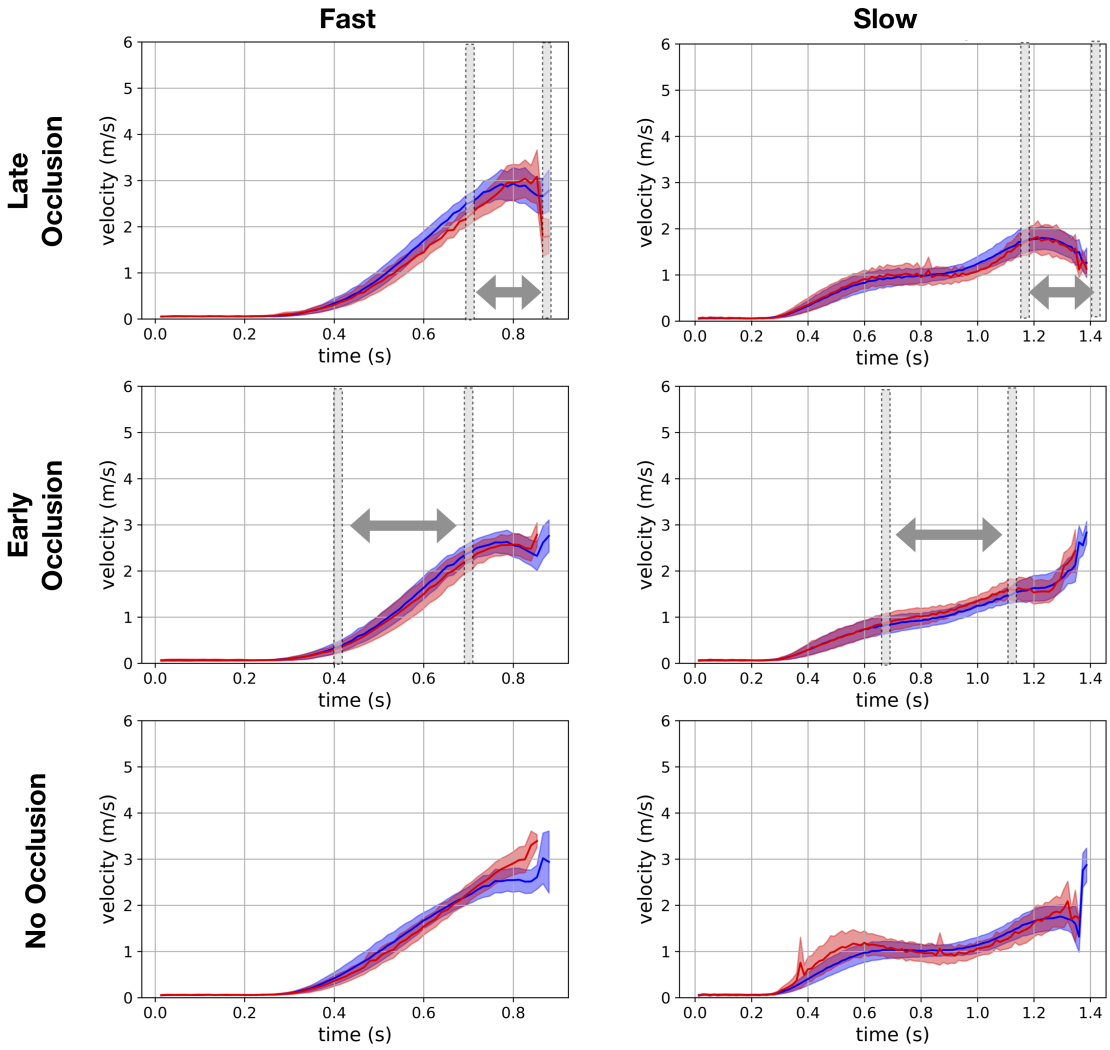
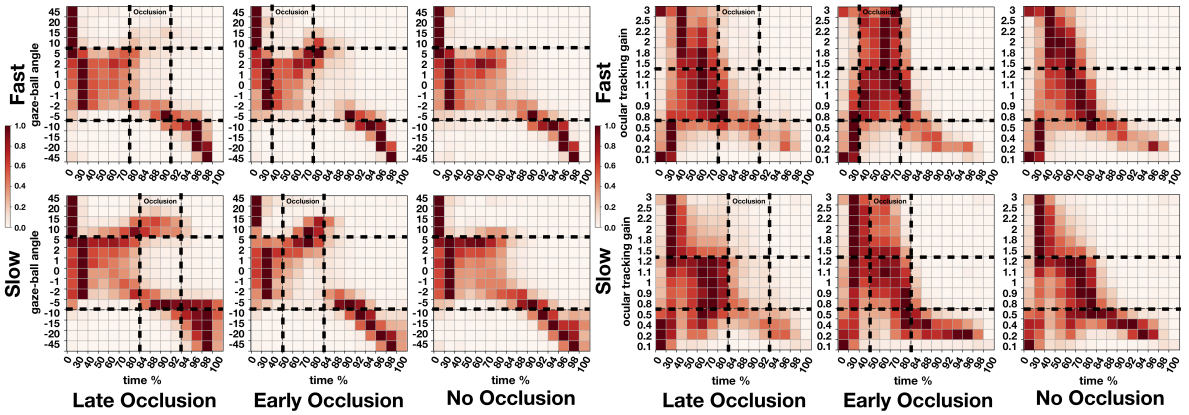


Figure 4.32: Hand velocity for all conditions

to bring about an interception was randomized, and did not differ between conditions. Visual inspection of Figure 4.32 shows that for the slow conditions, regardless of occlusion, velocity profiles appear to be similar across slow ball trajectories, regardless of the presence or timing of occlusion. Paddle velocity on slow trials reaches a plateau in the middle portion of the ball trajectory, until 150 ms before the catch, at which point the subjects make final adjustments to paddle position, resulting in a late peak in paddle velocity. In contrast, on fast ball trajectories, participants appear to reach a greater peak velocity earlier in the trial. However, the occlusion timing had no effect on the profile of absolute paddle velocity. This suggests that the subjects exploit a predictive strategy such that the motor planning remains almost unchanged during loss of visual information.

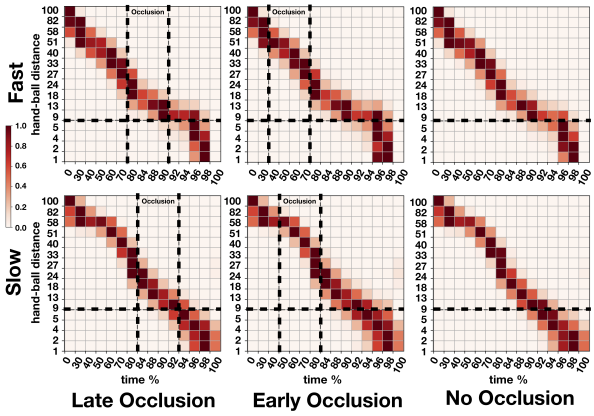
Estimated Reward Modules

Figure 4.33a & Figure 4.33b show the recovered reward values for the gaze agent using the method described in [118]. The top row of Figure 4.33a&Figure 4.33b correspond to the recovered reward modules are for the fast ball trajectories. These reward values are in agreement with the behavioral results presented earlier. In the early occlusion condition, subjects assigned higher reward values for the gaze leading the ball and also for the higher visual tracking gain. In contrast, for the late occlusion condition, reward values are similar to the no occlusion condition. Predictive eye movements are rewarded such that gaze falls behind the ball with lower tracking gain. For the slow condition, we also observe that for late occlusion, leading gaze positions are highly rewarded compared to the fast ball trajectory. This is consistent with the peaks observed in the top-right panel of Figure 4.30, which demonstrates that subjects tend to make predictive high velocity gaze movements to the estimated ball position a short duration into the future. The gaze reward modules capture detailed strategies of subjects eye movements and confirm that an RL framework with the estimated rewards shown in Figure 4.33a& Figure 4.33b can capture both predictive and on-line strategies driving the eye movements.



(a) Recovered gaze-to-ball angle reward for all conditions

(b) Recovered visual tracking gain reward for all conditions



(c) Recovered paddle distance reward for all conditions

Figure 4.33: Estimated gaze-ball angle, tracking gain and paddle distance reward modules for all subjects

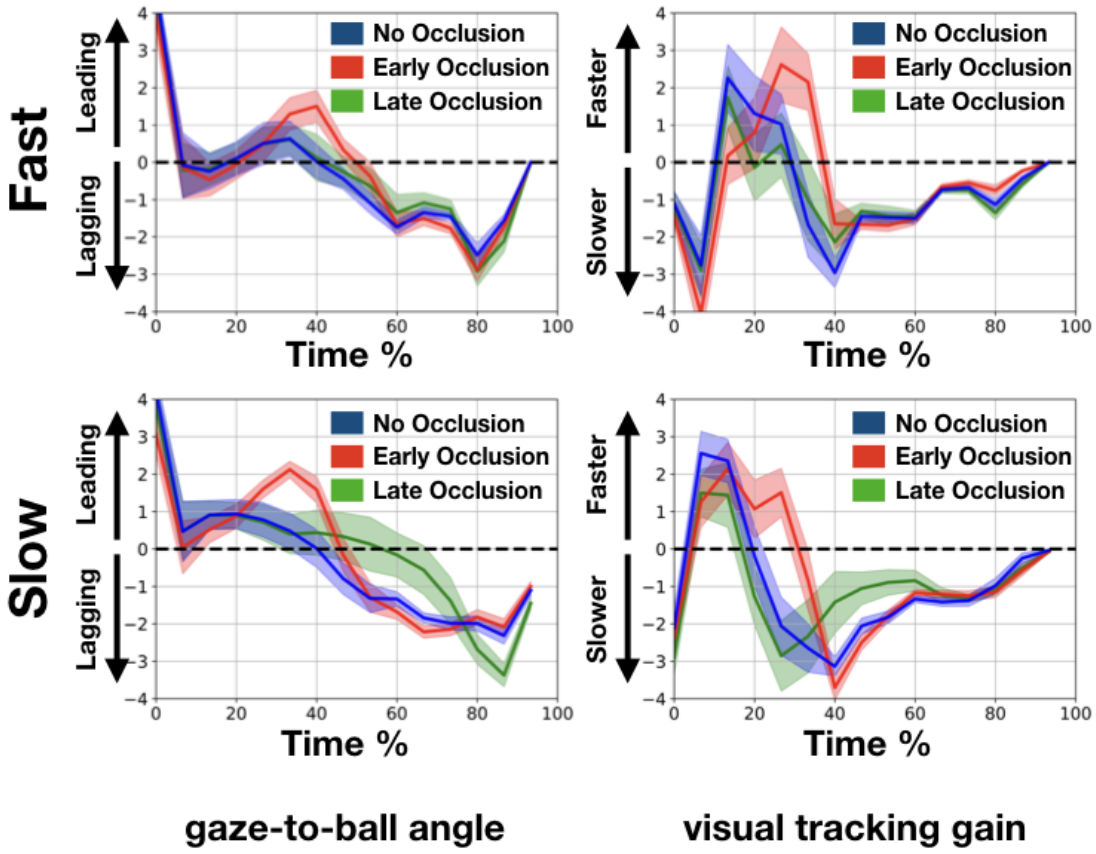


Figure 4.34: Recovered strategies of the subjects for visually tracking the ball. The shaded region represents the 95% confidence interval for between subject variance and the colors show the occlusion condition

Figure 4.33c shows the estimated reward modules for the paddle agent. The recovered reward for this submodule is aligned with our behavioral results in that there is no significant difference in strategies during different occlusion conditions. The model also captures slight differences between the slow and fast conditions, in that the slow condition reward module is more aligned with a reactive strategy. This suggests a slight change in hand movement strategy according the temporal demands of the task.

Transition Between Different Strategies for Individual Subjects

The reward modules presented in Figure 4.33a & Fig. 4.33b & Figure 4.33c are the results for the state-space observations of all 23 subjects. However, it is insightful to quantify the transition between predictive and on-line strategies for each subject. In order to do that, using IRL framework, we recovered individual subject's reward modules. In order to report the changes in the reward modules, we use a compact representation of the reward values that captures the variation between subjects as explained below.

If we horizontally divide the gaze reward modules, the top cells represent the reward values assigned to states in which the gaze leads and/or moves faster and the bottom cells where the gaze is lagging/moving slower than the ball. Considering Figure 4.33a & Figure 4.33b, the top cells are the states that produce predictive behavior compared to the bottom cells that the resulting behavior is more aligned with on-line strategy. If we multiply the top cell values by +1 and the bottom cell values by -1 and calculate the sum of reward values vertically, it provides a measure of dominant strategy at each time step. For example, for the gaze-to-ball reward module, if the reward value represents the state where the gaze leads the ball it will contribute to the summation as a positive value and if lagging, it would present a negative value. This strategy diagram is calculated for each subject and the mean and standard deviation from the mean is reported in Figure 4.34. This figure, not only shows the transition between predictive and on-line behavior according the timing of the occlusion, but also, it allows us to visualize the variability between subjects strategies at each time

step. For the gaze-to-ball angle module the red plot shows the leading vs lagging preference for the early occlusion condition.

4.4.5 Discussion

This study was designed to investigate the use of predictive strategies in a target interception task while the temporal demands of the task and the quality of visual information was systematically manipulated. A virtual ball catching experiment was designed to collect data from 23 subjects performing a visually guided target interception task. In order to be successful subjects were forced to use predictive strategies as the ball was occluded during early or late portions of the ball's parabolic trajectory. Subjects head, gaze and hand movements were recorded for further analysis. Results show, Subject's catching performance was significantly lower during early occlusion, suggesting the importance of early information of the ball trajectory. Also subject's catching performance was worse when attempting to catch a fast moving ball likely due to the need for relying on more predictive strategies. Subjects during the early occlusion consistently used predictive eye movements to where they estimated the ball would reappear after occlusion. This behavior suggests a transition to predictive eye movement strategy that facilitates visually tracking the ball immediately after the occlusion. For the late occlusion, only for the slow condition subjects performed a weaker predictive eye movement during occlusion, suggesting the effect of temporal demands of the task in changing the control strategy.

We also show that subjects modify/modulate the visual tracking gain during occlusion in order to track the ball right after occlusion. This change in visual tracking strategy was more apparent during the early occlusion compared to the late occlusion condition.

This predictive eye movement strategy could be accomplished through several mechanisms, including those that involve strong internal models of movement dynamics [21], or a hybrid approach that exploits information and learned mappings, rather than complex internal models or

extrapolations of the ball's movement through the 3D environment [112]. However, emerging evidence are in favor of strategies that don't require a model of projectile motion rather a memory-based extrapolation or a saccade to estimated future position could explain the behavior [61, 112]

Our behavioral results reveal a transition between predictive and on-line control strategies within a single interception task in response to changes in the temporal constraints and the timing of an occlusion within the ball trajectories. To capture this transition, an inverse reinforcement learning framework was used to visualize different strategies through an interpretation of recovered reward values associated with different states. The recovered rewards for gaze movements are consistent with the behavioral results. For example, during the early occlusion, off-line predictive gaze strategy along with higher visual tracking gain is highly rewarded. The results suggest that the gaze agent in order to produce behavior similar to human eye movements, need to modify its reward modules according to the condition. However, each reward module would produce behavior aligned with an on-line and/or predictive strategy.

Finally we characterized the transition between the two strategies by recovering the reward modules for individual subjects and also the between subject differences for each condition. Results help us report the between subject variability when switching from on-line to predictive strategy. Our proposed model allows us presenting the on-line vs prediction dichotomy in the form of spectrum of strategies rather than two separate mechanisms with no intermediate solutions in between.

Chapter 5

Discussion & Conclusion

Most of the classic and influential studies published in the field of vision science have been conducted in a constrained lab environment where the subjects could not perform an action as they would in a more natural context. Although these findings are fundamental to our understanding of the underlying mechanism of visual system, however it is important to investigate human behavior under a more naturalistic set of condition i.e., free movements of head and body.

Until recently, in order to study human visual system in a situation close to the real life, the experimenter would typically sacrifice the accuracy of stimulus properties or perform a tedious post-processing calculations. Therefore it was not straightforward to provide control on stimulus properties and the naturalness of the task. However, recent improvements in the temporal latency and spatial accuracy of VR displays make them suitable for scientific investigation. Using VR, we can present a parameterized visual stimulus and record behavior while subjects are allowed to move freely. Since the virtual content is pre-calculated and graphically presented on the display with respect to subject's head, the geometry and the physical properties of the stimulus can be precisely controlled by the experimenter. This dissertation presents a series of studies that leverage this new technology in the study of visual perception.

First, we presented a calibration method to reduce the spatial and temporal errors in an SMI eye tracking integration. Our post-hoc calibration method reduced the static angular error in the periphery and used a dynamic re-calibration to compensate for small physical shifts of the head mounted display relative to subject's face. This method improves the eye tracking accuracy for research purposes without extending the re-calibration sessions that happens in between data collection [30, 48, 121]. These results along with other findings are presented to the vision science community as the use of VR systems in vision research is rapidly growing.

We then presented several studies that leveraged the improved eye tracking signal to investigate the underlying predictive strategies driving our gaze behavior and hand movements during a visually guided target interception. We chose this paradigm because it provides the opportunity to study different aspects of visual processing system. For example, when we attempt to catch a ball moving through our field of view, certain information regarding the kinematics of the ball trajectory needs to be perceived/calculated by our visual system, the perceived motion and its properties are then mapped to a motor command to be executed by the muscles. In the literature it was shown that in order to be successful, the appropriate movement has to be programmed, and consequence of that movement needs to be taken into account in a predictive way. However, here we proposed a model that showed how this prediction can be implemented via recursive model that was originally inspired by the mechanism of our memory (LSTM-RNN). We showed a temporal mapping between visual information about the ball in the past to the motor output in the distant future can reproduce the human behavior with high accuracy.

In our first experiment, we showed that the action is generally guided through an online strategy unless there is a need for prediction. We forced our subjects to the predictive mode of control by making the ball disappear for a fixed 500 ms duration midway through its parabolic trajectory. We showed that the quality of off-line predictive eye movements during the blank by comparing subjects gaze vs ball position/velocity. Results showed that subjects gaze trajectory was well matched with the curvature of the ball trajectory during blank. These results provide a more detail

understanding of predictive hand and gaze movements in a more naturalistic condition.

Subjects' catching performance was worse when they had shorter post-blank duration. For example, for the two pairs of conditions within which the flight time was the same, subjects missed the ball more when the post-blank duration was shorter. We also found that the accuracy of predictive eye movements was correlated with the accuracy of the hand placement. This suggests that the predictive strategies used by the subjects could result from a common representation in the brain as it was suggested in previous studies before [72]

In the follow up study we focused on the possible mechanisms that explain prediction. We used the collected data to create a computational model that characterizes the relationship between information received and processed by our sensory system with the produced motor output in a predictive way. We used LSTM-RNNs to capture the predictive mechanisms driving our eye-hand movements by training the model on the subjects data. This computational model was specifically designed based on the notion of hybrid control model proposed by Zhao et al[15] such that it only sees the world through a parsimonous list of sensory input and predicts the motor command 300 ms into the future. The model was able to predict into the future with reasonable accuracy, and serves as proof of concept that prediction does not explicitly require an internal model of the world dynamics. We also found that when we increased the temporal duration of the window over which visual information was integrated, the model's prediction accuracy did not increase significantly. However, it made the model more robust when dealing with sensory perturbation. The model with a longer integration window produced less error when the input sensory information was removed.

In addition, we investigated the role of individual sources of information. Two identical models were trained for two distinct groups of subjects based on their catching performance. Following training, information sources were selectively removed, and the degradation to model performance considered an estimate of the source of the ablated sources of sensory information. The model trained on data from the more successful group produced less error during the ablation study. This suggests that the more successful group might have relied on a wider range of input sensory infor-

mation so that removing one source of information does not affect their performance significantly.

In the final study we focused on two important aspects of a successful visually guided hand movement strategy. First we compared the quality of information during early and late portion of a ball trajectory by systematically placing an occluder in front of the subjects. Results showed that subjects missed the ball the most in the early occlusion condition suggesting the importance of early visual information studied in the literature [85, 86].

Second, we investigated the effect of temporal demands of the task on switching between online and predictive strategy. We used fast vs slow trajectories to investigate the switch between online and predictive strategies. We observed that during late occlusion condition where the velocity of the ball is high and makes it nearly impossible to visually track the ball, subjects increased their visual tracking gain only in the slow condition. This confirms our hypothesis that beside target occlusion, temporal demands of the task made the subject actively adapt their visual tracking strategy accordingly.

As a general conclusion, the author finds the following items as the broader contribution of this dissertation to the field. Extending the study of human visual system to a new environment with systematic considerations about experimental apparatus provided a new opportunity to answer important questions in a close to real life setup. This will help us in generalizing the results to a real world condition.

Prediction is a key element of our visual system and this work provided strong observations of its presence in a visually guided action. Humans use predictive strategies when they would not be successful using online control strategies because of the temporal demands of the task. In addition, we showed that predictive strategies do not necessarily require an internal model of the world kinematics and it could be explained using a temporal mapping of input sensory information to an action in future. This is rooted in theories of visual perception where the available information for our visual system along with proprioceptive signals are proven to be sufficient to explain short-lived predictive strategies [7].

Finally, the dichotomy of predictive vs on-line control can be explained using a single model that captures the representation of these two types of behavior. Our results show that in the course of a single action, humans switch between these strategies based on the quality of available information and timing of the task in hand. By fitting an inverse reinforcement learning model to subjects data, we showed that using an action-reward framework for a gaze and paddle agent, we can explain their on-line control and also predictive strategies. This suggests the use of a similar rewarding mechanism in our sensory system that drives our actions shown in other contexts [122, 123].

Chapter 6

Future Work

In this section, we provide possible directions for the future of this research project based on the limitations and constraints of the current study. The limitations of the current work could be categorized into two major classes.

First, the constraints related to the systems and the technology used in this study. In order to investigate predictive eye movements we utilized an eye tracking system built into a VR head mounted display. Any eye tracking system reports the position of the gaze in the scene where the user is foveating. Although our visual acuity, motion perception and motion discrimination is highest at the fovea, however, peripheral vision could also be helping the subjects in order to estimate the direction of the ball especially toward the end of the trial. Therefore, in the current study we could not monitor or characterize the use of peripheral visual information. It is reasonable to hypothesize that our visual processing and motor planning system has developed some sort of mapping between the motion sensed in the periphery and the kinematics of the ball during the final interception phase. Even though the perceived motion direction could be inaccurate [124, 125]. One way we could address this issue is to find the correlation between subjects' peripheral motion perception and their performance in the task. Finding a correlation between these two metrics

could suggest the influence of peripheral motion processing on the performance during a visually guided action.

Second, our proposed computational model assumed that the knowledge of visual and non-visual information from the past is not subject to temporal decay. Although the LSTM-RNN models are inspired by our brain and the way it can utilize temporal information in the past, but we suggest the use of a decaying mechanism in the model in order to realistically map the information from the past to future. Furthermore, we could also include the effect of past experience in the model by inviting the subjects to participate in the experiment for multiple days and report the change in models activation pattern, weight distributions in the hope for finding learning and experience effects.

We also propose three directions that this research could be extended upon. First, instead of using a virtual reality setup to record human behavior one could extend this study in an outdoor sport venue where the subjects will be experiencing a more realistic condition and at the same time avoid limitations of a VR system such as small FOV and end-to-end latency. It is important to note that, recording subject's motion and eye tracking would be a challenge as shown in previous studies [72]. Recently, outdoor motion sensing technology with inertial measurement units has proven to provide promising results. Also, portable eye tracking systems are now available with lower price and higher accuracy. Especially the user interface of new eye tracking devices allow recalibration and post-hoc corrections of the eye tracking data in a user friendly fashion. Furthermore, the advancements in semantic segmentation techniques allow us to localize the objects in the scene image and hence calculate object positions with respect to subject's eye position [126]. One could setup a similar task in a real sport venue and use natural perturbation that might occur to a tennis player, i.e. the target getting occluded by the player. Therefore, the advances in portable eye tracking hardware and software along with state of the art image processing techniques presents a great future possibility for this study so that, not only the subject is not limited to the field of view of a VR display or system latency - that can possibly create biases in their head movements

- but also the experimenter would face less processing and computational challenges in order to calculate the properties of the stimulus in the scene image.

Second, with a more general definition of the task, such as hitting the ball toward a predefined direction, more complicated tasks could be investigated and the results would provide a means for comparison with the current results. We predict that the hand movement pattern for hitting a ball toward a certain direction would be different than catching, hence it forces the subjects to exploit different strategies during planning and execution. By modifying the task, we expect to observe a modified pattern of hand movements in order to redirect the ball into a certain direction. In order to do that, subjects would ideally need to plan ahead of time and possibly assign part of their attention to the final destination rather than solely tracking the moving ball. Recent studies used a similar paradigm and reported that for a baseball out fielder, in order to run or look back and track the ball, subjects demonstrate an optimum decision making strategy that is well explained by the ratio of system to observation noise and the ratio between reaction time and task duration [84]. Our hypothesis is that subjects would switch their visual tracking strategy to the competing tasks according to the temporal demands of the task or according to their ability to predict the ball trajectory.

Third, since the data set provided in this study includes the right and left eye scene images, one possible future direction to this research would be to use input scene images (monocular or binocular) to generate motor response using computer vision and machine learning techniques. The recent growth of deep learning methods has provided promising results such that a deep network will be able to extract features from the images probably in a similar way our visual system does. However, there will be a debate on whether the representations captured by a deep network is similar to the human visual system or there are differences. This could also be accompanied with a computational model of foveated vision so that the input scene image goes through the distortion and transformations similar to what our visual system perceives. Recent studies show promising results when an ideal observer model is compared with human observer through a computational

model that captures the characteristics of visual processing pipeline for motion perception [127].

Finally, this study could also be designed in order to focus on expert athletes and provide a systematic measurement of their visual-motor performance. There are many sports that would benefit from the findings of this study and could implement the setup as a coaching tool to quantify their performance metrics. For instance it was shown that when comparing expert vs. novice volleyball players, the pattern of saccadic and fixation eye movements were significantly different, suggesting a possible coaching regime for improvements in athletes performance [128].

Bibliography

[1] Jan Koenderink, Whitman Richards, and Andrea J van Doorn. Space-time disarray and visual awareness. i-Perception, 3(3):159–165, 2012.

[2] William H Warren. Does this computational theory solve the right problem? marr, gibson, and the goal of vision. Perception, 41(9):1053–1060, 2012.

[3] William H Warren Jr, Bruce A Kay, Wendy D Zosh, Andrew P Duchon, and Stephanie Sahuc. Optic flow is used to control human walking. Nature neuroscience, 4(2):213, 2001.

[4] William H Warren. How do animals get about by vision? Visually controlled locomotion and orientation after 50 years. 100(0 0):277–281, 2013.

[5] William H Warren Jr. Visually controlled locomotion: 40 years later. Ecological Psychology, 10(3-4):177–219, 1998.

[6] Myrka Zago and Francesco Lacquaniti. Visual perception and interception of falling objects: a review of evidence for an internal model of gravity. Journal of Neural Engineering, 2(3):S198, 2005.

[7] James J. Gibson. The Ecological Approach to the Visual Perception of Pictures. Leonardo, 11(3):227–235, 1978. ISSN 0024094X.

[8] Richard L Gregory. Forty years on: Kenneth craik's the nature of explanation (1943), 1983.

[9] James J. Gibson. The Ecological Approach to Visual Perception. Houghton Mifflin, Boston, 1979. ISBN 0898599598.

[10] University of New South Wales. Neuroscientist discovers hidden region in the human brain, 2018.

[11] David Marr. Visual Information Processing: The Structure and Creation of Visual Representation. doi: 10.1098/rstb.2010.0098.

[12] Jan Koenderink. Vision as a User Interface. Proceedings of SPIE-IS&T Electronic Imaging, 7865:13, 2011.

[13] Jan Koenderink. Gestalts as ecological templates. pages 1–17, 2003.

[14] S. Lappin, Joseph. Inferential and Ecological theories of Visual Perception. (September), 2016.

[15] William H Warren. On-line and model-based approaches to the visual control of action.

[16] GR Barnes and PT Asselman. The mechanism of prediction in human smooth pursuit eye movements. The Journal of physiology, 439(1):439–461, 1991.

[17] Richard L Gregory. Forty years on: Kenneth craik's the nature of explanation (1943), 1983.

[18] Daniel E Rivera. Internal model control: A comprehensive view. Arizona State University, Tempe, Arizona, pages 85287–6006, 1999.

[19] Paul R Davidson and Daniel M Wolpert. Widespread access to predictive models in the motor system: a short review. Journal of Neural Engineering, 2(3):S313, 2005.

[20] Emo Todorov. Neural control of movement: A computational perspective.

[21] Daniel M Wolpert and J.Randall Flanagan. Motor prediction. Current Biology, 11(18): R729 – R732, 2001.

[22] Myrka Zago and Francesco Lacquaniti. Visual perception and interception of falling objects: a review of evidence for an internal model of gravity. Journal of Neural Engineering, 2(3):S198, 2005.

[23] Mitsuo Kawato. Internal models for motor control and trajectory planning. Current Opinion in Neurobiology, 9(6):718–727, 1999.

[24] D M Wolpert and Z Ghahramani. Computational principles of movement neuroscience. Nature Neuroscience, (november):1212–7, November 2000.

[25] Garey H Noritz, Nancy A Murphy, et al. Motor delays: early identification and evaluation. Pediatrics, pages peds–2013, 2013.

[26] Steven W. Keele and Michael I. Posner. Processing of visual feedback in rapid movements. Journal of Experimental Psychology, 77(1):155–158, 1968.

[27] Dave Carlson Website. Eye anatomy, 2018.

[28] Roy H Steinberg, Miriam Reid, and Paula L Lacy. The distribution of rods and cones in the retina of the cat (felis domesticus). Journal of Comparative Neurology, 148(2):229–248, 1973.

[29] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. Eye tracking: A comprehensive guide to methods and measures. OUP Oxford, 2011.

[30] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. Eye tracker data quality. In ETRA '14, page 45, New York, New York, USA, 2012. ACM Press.

[31] AL Yarbus. Eye movements and vision. 1967. New York, 1967.

[32] Diederick C Niehorster, Tim HW Cornelissen, Kenneth Holmqvist, Ignace TC Hooge, and Roy S Hessels. What to expect from your remote eye-tracker when participants are unrestrained. Behavior research methods, 50(1):213–227, 2018.

[33] Miriam Spering, Alexander C Schütz, Doris I Braun, and Karl R Gegenfurtner. Keep your eyes on the ball: smooth pursuit eye movements enhance prediction of visual motion. Journal of Neurophysiology, 105(4):1756–67, April 2011. ISSN 1522-1598.

[34] Mary Hayhoe, Travis McKinney, Kelly Chajka, and Jeff Pelz. Predictive eye movements in natural vision. Experimental Brain Research, 217(1):125–36, mar 2012. ISSN 1432-1106.

[35] Brett R Fajen and Michael C Devaney. Learning to control collisions: the role of perceptual attunement and action boundaries. Journal of experimental psychology. Human perception and performance, 32(2):300–13, apr 2006. ISSN 0096-1523.

[36] Yunfeng Zhang and Anthony J Hornof. Easy post-hoc spatial recalibration of eye tracking data. In ETRA '14, pages 95–98, New York, New York, USA, 2014. ACM Press.

[37] Joseph J LaViola Jr, Daniel Acevedo Feliz, Daniel F Keefe, and Robert C Zeleznik. Hands-free multi-scale navigation in virtual environments. In Proceedings of the 2001 Symposium on Interactive 3D Graphics, pages 9–15. ACM, 2001.

[38] Leigh A Mrotek and John F Soechting. Target interception: hand–eye coordination and strategies. Journal of Neuroscience, 27(27):7297–7309, 2007.

[39] Liana E Brown, Elizabeth T Wilson, Melvyn A Goodale, and Paul L Gribble. Motor force field learning influences visual processing of target motion. Journal of Neuroscience, 27 (37):9975–9983, 2007.

[40] Laurel A Issen and David C Knill. Decoupling eye and hand movement control: visual short-term memory influences reach planning more than saccade planning. Journal of vision, 12(1):3–3, 2012.

[41] John F Soechting, John Z Juveli, and Hrishikesh M Rao. Models for the extrapolation of target motion for manual interception. Journal of Neurophysiology, 102(3):1491–1502, 2009.

[42] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. Trends in cognitive sciences, 9(4):188–194, 2005.

[43] John W Krakauer, Zachary M Pine, Maria-Felice Ghilardi, and Claude Ghez. Learning of visuomotor transformations for vectorial planning of reaching trajectories. Journal of Neuroscience, 20(23):8916–8924, 2000.

[44] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision, 87(1):4–27, 2010.

[45] Sebastian Friston and Anthony Steed. Measuring latency in virtual environments. IEEE Transactions on Visualization and Computer Graphics, 20(4):616–625, 2014.

[46] Katerina Mania, Bernard D Adelstein, Stephen R Ellis, and Michael I Hill. Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity. In Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization, pages 39–47. ACM, 2004.

[47] Bernard D Adelstein, Thomas G Lee, and Stephen R Ellis. Head tracking latency in virtual environments: psychophysics and a model. In Proceedings of the Human Factors and

Ergonomics Society Annual Meeting, volume 47, pages 2083–2087. SAGE Publications Sage CA: Los Angeles, CA, 2003.

[48] Kamran Binaee, Gabriel Diaz, Jeff Pelz, and Flip Phillips. Binocular Eye tracking Calibration During a Virtual Ball Catching task using Head Mounted Display. In Proceedings of the ACM Symposium on Applied Perception, pages 15–18, 2016.

[49] Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. Machine learning: An Artificial Intelligence Approach. Springer Science & Business Media, 2013.

[50] John McCarthy and Edward A Feigenbaum. In memoriam: Arthur samuel: Pioneer in machine learning. AI Magazine, 11(3):10–10, 1990.

[51] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[52] Yoshua Bengio, Ian J Goodfellow, and Aaron Courville. Deep learning. Nature, 521:436–444, 2015.

[53] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction, volume 1. MIT press Cambridge, 1998.

[54] Christopher JCH Watkins and Peter Dayan. Q-learning. Machine learning, 8(3-4):279–292, 1992.

[55] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In Icml, pages 663–670, 2000.

[56] Ryan Arzamarski, Steven J. Harrison, Alen Hajnal, and Claire F. Michaels. Lateral ball interception: hand movements during linear ball trajectories. Experimental Brain Research, 177(3):312–323, 2007.

[57] Siavash Vaziri and Charles E Connor. Representation of gravity-aligned scene structure in ventral pathway visual cortex. Current Biology, 26(6):766–774, 2016.

[58] Alen Hajnal, Michael Grocki, David M Jacobs, Frank TJM Zaal, and Claire F Michaels. Mode transition and change in variable use in perceptual learning. Ecological Psychology, 18(2):67–91, 2006.

[59] David L. Mann, Wayne Spratford, and Bruce Abernethy. The Head Tracks and Gaze Predicts: How the World's Best Batters Hit a Ball. PLoS ONE, 8(3):e58289, mar 2013. ISSN 1932-6203.

[60] Gabriel Jacob Diaz, J Cooper, C Rothkopf, and M Hayhoe. Saccades to future ball location reveal memory-based prediction in a virtual-reality interception task. Journal of Vision, 13 (1):20–20, January 2013.

[61] Gabriel Diaz, Joseph Cooper, Mary Hayhoe, and Phil Trans R Soc B. Memory and prediction in natural gaze control. Philosophical Transations of the Royal Society B: Biological Sciences, 368(1628), 2013.

[62] M F Land and P McLeod. From eye movements to actions: how batsmen hit the ball. Nature Neuroscience, 3(12):1340–5, dec 2000. ISSN 1097-6256.

[63] Benedetta Cesqui, Maura Mezzetti, Francesco Lacquaniti, and Andrea D'Avella. Gaze Behavior in One-Handed Catching and Its Relation with Interceptive Performance: What the Eyes Can't Tell. Plos One, 10(3):e0119445, 2015. doi: 10.1371/journal.pone.0119445.

[64] Ramesh Raskar, Hideaki Nii, Bert Dedecker, Yuki Hashimoto, Jay Summet, Dylan Moore, Yong Zhao, Jonathan Westhues, Paul Dietz, John Barnwell, et al. Prakash: lighting aware motion capture using photosensing markers and multiplexed illuminators. In ACM Transactions on Graphics (TOG), volume 26, page 36. ACM, 2007.

[65] Shoichi Shimizu and Hironobu Fujiyoshi. Acquisition of 3D gaze information from eyeball movements using inside-out camera. In Proceedings of the 2nd Augmented Human International Conference - AH '11, pages 1–7, New York, New York, USA, 2011. ACM Press.

[66] Pieter Blignaut and Daniël Wium. The effect of mapping function on the accuracy of a video-based eye tracker. In Eye Tracking South Africa, pages 39–46, Cape Town, 2013. ACM Press.

[67] Andrew T. Duchowski Rui I. Wang, Brandon Pelfrey and Donald H. House. Online 3d gaze localization on stereoscopic displays. ACM Trans. Appl. Percept, 1(3):11, 2014.

[68] Wechsler H. Duchowski A. T. Ji, Q. and M. Flickner. Special issue on eye detection and tracking. Computer Vision and Image Understanding, 1(98), 2005.

[69] Adrian Haffegee, Vassil Alexandrov, and Russell Barrow. Eye tracking and gaze vector calculation within immersive virtual environments. In VRST 2007, page 225, Newport Beach, CA, 2007. ACM Press.

[70] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395, 1981.

[71] P Possidente, Flip Phillips, J S Matthis, and Gabriel Jacob Diaz. Anticipation of sabre fencing attacks. Journal of Vision, 11(11):957–957, September 2011.

[72] Mary M Hayhoe, Travis McKinney, Kelly Chajka, and Jeff B Pelz. Predictive eye movements in natural vision. Experimental Brain Research, 217(1):125–136, 2012.

[73] Jolande Fooken, Sang-Hoon Yeo, DInesh K. Pai, and Miriam Spering. Eye movement

accuracy determines natural interception strategies. Journal of Vision, 16(14):1–15, 2016. doi: 10.1167/16.14.1.doi.

[74] Gilles Montagne. Prospective control in sport. International Journal of Sport Psychology, 36(2):127–150, 2005.

[75] Graham R Barnes and C J Sue Collins. The influence of cues and stimulus history on the non-linear frequency characteristics of the pursuit response to randomized target motion. Experimental Brain Research., 212(2):225–40, July 2011. ISSN 1432-1106.

[76] Mary Hayhoe and Dana Ballard. Modeling task control of eye movements. Current Biology, 24(13):R622–R628, 2014.

[77] Liesbeth I N Mazyn, Geert J P Savelsbergh, Gilles Montagne, and Matthieu Lenoir. Planning and on-line control of catching as a function of perceptual-motor constraints. Acta psychologica, 126(1):59–78, sep 2007.

[78] Aldo Faisal and Daniel M Wolpert. Near optimal combination of sensory and motor uncertainty in time during a naturalistic perception-action task. Journal of Neurophysiology, 101 (4):1901–1912, 2009. ISSN 0022-3077. doi: 10.1152/jn.90974.2008.

[79] R. H. Sharp and H. T.A. Whiting. Exposure and occluded duration effects in a ball-catching skill. Journal of Motor Behavior, 1974. ISSN 19401027. doi: 10.1080/00222895.1974.10734990.

[80] Alain Zuur, Elena N Ieno, Neil Walker, Anatoly A Saveliev, and Graham M Smith. Mixed effects models and extensions in ecology with R. Springer Science & Business Media, 2009.

[81] George EP Box and David R Cox. An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological), 26(2):211–243, 1964.

[82] Travis McKinney, Kelly Chajka, and Mary Hayhoe. Pro-active gaze control in squash. Journal of Vision, 8(6):111, 2008.

[83] Samuel Tuhkanen, Jami Pekkanen, Paavo Rinkkala, Callum Mole, Richard M Wilkie, and Otto Lappi. Humans use predictive gaze strategies to target waypoints for steering. Scientific Reports, 9(1):8344, 2019.

[84] Boris Belousov, Gerhard Neumann, Constantin A Rothkopf, and Jan R Peters. Catching heuristics are optimal control policies. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 1426–1434. Curran Associates, Inc., 2016.

[85] Joan López-Moliner, Eli Brenner, Stefan Louw, and Jeroen BJ Smeets. Catching a gently thrown ball. Experimental Brain Research, 206(4):409–417, 2010.

[86] Cristina de la Malla and Joan Lopez-Moliner. Predictive Plus Online Visual Information Optimizes Temporal Precision in Interception. Journal of experimental psychology: Human perception and performance, 41(5):1271–1280, 2015.

[87] Kielan Yarrow, Peter Brown, and John W Krakauer. Inside the brain of an elite athlete: the neural processes that support high achievement in sports. Nature reviews. Neuroscience, 10 (8):585–96, aug 2009. ISSN 1471-0048.

[88] A T Bahill, D Adler, and L Stark. Most naturally occurring human saccades have magnitudes of 15 degrees or less. Investigative Ophthalmology, 14:468–469, 1975. ISSN 0020-9988.

[89] Joost C. Dessing, Frédéric P. Rey, and Peter J. Beek. Gaze fixation improves the stability of expert juggling. Experimental Brain Research, 216(4):635–644, 2012. ISSN 00144819. doi: 10.1007/s00221-011-2967-6.

[90] Raoul Huys, Andreas Daffertshofer, and Peter J. Beek. Multiple time scales and subsystem embedding in the learning of juggling, 2004. ISSN 01679457.

[91] Sergio T. Rodrigues, Joan N. Vickers, and A. Mark Williams. Head, eye and arm coordination in table tennis. Journal of Sports Sciences, 20(3):187–200, 2002. ISSN 02640414. doi: 10.1080/026404102317284754.

[92] Hubert Ripoll and Philippe Fleurance. What does keeping one's eye on the ball mean? Ergonomics, 31(11):1647–1654, 1988.

[93] Robert N. Singer, A. Mark Williams, Shane G. Frehlich, Christopher M. Janelle, Steven J. Radlo, Douglas A. Barba, and Lester J. Bouchard. New frontiers in visual search: An exploratory study in live tennis situations. Research Quarterly for Exercise and Sport, 69 (3):290–296, 1998.

[94] Huaiyong Zhao and William H Warren. Intercepting a moving target: On-line or model-based control? Journal of vision, 17(5):12–12, 2017.

[95] DN David N Lee. A theory of visual control of braking based on information about time to collision. Perception, 5:437–459, 1976.

[96] R.J. Bootsma and P.C.W. van Wieringen. Timing an attacking forehand drive in table tennis. Journal of Experimental Psychology: Human Perception and Performance, 16(1):21–29, 1990.

[97] G. J.P. Savelsbergh, H. T.A. Whiting, J. R. Pijpers, and A. A.M. van Santvoord. The visual guidance of catching. Experimental Brain Research, 93(1):148–156, 1993. ISSN 00144819.

[98] Jr Tresilian. Visually timed action: time-out for 'tau'? Trends in cognitive sciences, 3(8): 301–310, aug 1999.

[99] Joan López-Moliner, Hans Supèr, and Matthias Keil. The time course of estimating time-to-contact: switching between sources of information. Vision research, (September):5–10, sep 2013.

[100] Cristina de la Malla and Joan López-Moliner. Predictive plus online visual information optimizes temporal precision in interception. Journal of experimental psychology: human perception and performance, 41(5):1271, 2015.

[101] Myrka Zago, Joseph McIntyre, Patrice Senot, and Francesco Lacquaniti. Visuo-motor coordination and internal models for object interception. Experimental Brain Research., 192 (4):571–604, feb 2009.

[102] Damian G Stephen, Nigel Stepp, James A Dixon, and MT Turvey. Strong anticipation: Sensitivity to long-range correlations in synchronization behavior. Physica A: Statistical Mechanics and its Applications, 387(21):5271–5278, 2008.

[103] Zainy MH Almurad, Clément Roume, and Didier Delignières. Complexity matching in side-by-side walking. Human Movement Science, 54:125–136, 2017.

[104] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, pages 6645–6649. IEEE, 2013.

[105] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[106] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In Interspeech, pages 194–197, 2012.

[107] Hamid Khodabandehlou and Mohammad Sami Fadali. Echo state versus wavelet neural networks: Comparison and application to nonlinear system identification. IFAC-PapersOnLine, 50(1):2800–2805, 2017.

[108] Hamid Khodabandehlou and Mohammad Sami Fadali. Networked control of unmanned vehicle using wavelet-based generalized predictive controller. In 2016 International Joint Conference on Neural Networks (IJCNN), pages 5226–5233. IEEE, 2016.

[109] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[110] James E Cutting. Potency, and contextual use of different in_fi) rmation about depth. Perception of space and motion, page 69, 1995.

[111] Laurel A Issen and David C Knill. Decoupling Eye and Hand Movement Control: Visual Short-term Memory Influences Reach Planning More than Saccade Planning. Journal of vision, 12(1):1–13, jan 2012. ISSN 1534-7362.

[112] Kamran Binaee, Anna Starynska, Jeff B Pelz, Christopher Kanan, and Gabriel Jacob Diaz. Characterizing the temporal dynamics of information in visually guided predictive control using lstm recurrent neural networks. arXiv preprint arXiv:1805.05946, 2018.

[113] Joan Lopez-Moliner and Eli Brenner. Flexible timing of eye movements when catching a ball. Journal of vision, 16(5):13–13, 2016.

[114] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. arXiv preprint arXiv:1710.04615, 2017.

[115] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research, 32(11):1238–1274, 2013.

[116] Matthew H Tong, Oran Zohar, and Mary M Hayhoe. Control of gaze while walking: Task structure, reward, and uncertainty. Journal of vision, 17(1):28–28, 2017.

[117] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In AAAI, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[118] Matthew Alger. Inverse reinforcement learning, 2016.

[119] Jamesj. Oculus discussion forums, Sep 2014. URL https://forums.oculusvr.com/developer/discussion/14796/what-is-the-angular-resolution-of-the-dk2.

[120] Masayuki Iwasaki and H Inomata. Relation between superficial capillaries and foveal structures in the human retina. Investigative Ophthalmology & Visual Science, 27(12):1698–1705, 1986.

[121] Jason S Babcock and Jeff B Pelz. Building a lightweight eyetracking headgear. In Proceedings of the 2004 symposium on Eye tracking research & applications, pages 109–114. ACM, 2004.

[122] Ruohan Zhang, Shun Zhang, Matthew H. Tong, Yuchen Cui, Constantin A. Rothkopf, Dana H. Ballard, and Mary M. Hayhoe. Modeling sensory-motor decisions in natural behavior. PLOS Computational Biology, 14(10):1–22, 10 2018.

[123] Jean-Jacques Orban de Xivry, Sébastien Coppe, Gunnar Blohm, and Philippe Lefevre. Kalman filtering naturally accounts for visually guided and predictive smooth pursuit dynamics. Journal of Neuroscience, 33(44):17301–17313, 2013.

[124] W Monaco, J Kalb, and Chris A Johnson. Motion detection in the far peripheral visual field. Army Research Laboratory Report ARL-MR-06, 2007.

[125] Duje Tadin, Jeffrey B Nyquist, Kelly E Lusk, Anne L Corn, and Joseph S Lappin. Peripheral vision of youths with low vision: motion perception, crowding, and visual search. Investigative ophthalmology & visual science, 53(9):5860–5868, 2012.

[126] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

[127] Benjamin M Chin and Johannes Burge. Predicting the partition of behavioral variability in speed perception with naturalistic stimuli. bioRxiv, page 601161, 2019.

[128] A Piras, R Lobietti, and S Squatrito. A study of saccadic eye movement dynamics in volleyball: comparison between athletes and non-athletes. Journal of Sports Medicine and Physical Fitness, 50(1):99, 2010.